University of Canterbury
Department of Mathematics and Statistics

Nearest Neighbour Imputation and Variance Estimation Methods

A thesis submitted in
partial fulfilment
of the requirements of
the Degree for
PhD in Statistics
at the
University of Canterbury
by

Murthy N. Mittinty

Supervisor: Dr Easaw Chacko

2004

To

My Teachers

Acknowledgements

I wish to express my sincere appreciation to all those people who have prayed, wished, and supported me throughout my education.

I am highly indebted to my supervisors Dr. Easaw Chacko, and Mr. Richard Penny for their constant encouragement, support, guidance and patience in going through my work. I am also thankful to Dr. Monir Hossain who had also been my co-supervisor. I also acknowledge the support that I received from faculty and administrative staff of the Department of Mathematics and Statistics at the University of Canterbury. My thanks are also to the Physical Science Library staff who helped me in getting the articles and inter loans on time. I thank the staff at Ilam village and also the International Student Support who helped me in every aspect of my stay in New Zealand.

I would like to thank the Ministry of Foreign Affairs and Trade, New Zealand, for the NZAID study awards, and the University of Canterbury (UOC) for the UOC Doctoral award, without which this study would had been impossible. Last but not the least I thank my parents my sister and brother-in-law and their son for their support and encouragement.

Papers Resulting from this Thesis

Published papers

- Murthy M. N., Chacko E., Penny Richard and Hossain M.M.(2003) Multivariate nearest neighbourhood method of imputation. *Statistics in Transition*, 6(1):55-66.

- Data reduction by principal components and Graphical modelling (2003), Presented at the 54^{th} ISI, Meetings, Berlin (Published in Conference Proceedings).

- Murthy M.N., Chacko E. (2005) Imputation by Propensity Matching. *In American Statistical Association, Proceedings of the Survey Research Methods*, pages[CD-ROM] pp4022-4028.

Conference Presentations

- Multivariate nearest neighbourhood method of imputation. (2002), Presented at Baltic Nordic conference on Survey Sampling, August 17-23, 2002 in Ammarns, Sweden.

- Data reduction by principal components and Graphical modelling (2003), Presented at the 54^{th} ISI, Meetings, Berlin .

- Imputation by Propensity Matching (2004), Paper presented at the Joint Statistical Meetings Toronto.

Paper in Progress

- Bootstrap Variance Estimation for Two Stage Cluster Designs with Imputed Data. Paper submitted at the 55th ISI Sydney conference. To be held in April 2005.

Imputation and Variance Estimation Methods

Murthy. N. Mittinty[1]

[1]Department of Mathematics and Statistics, University of Canterbury, New Zealand, email: nmi13@student.canterbury.ac.nz

Abstract

In large-scale surveys, nonresponse is a common phenomenon. This nonresponse can be of two types; unit and item nonresponse. In this thesis we deal with item nonresponse as other responses from the survey unit can be used for adjustment. Usually nonresponse adjustment is carried out in one of three ways; weighting, imputation and no adjustments. Imputation is the most commonly used adjustment method, either as single imputation or multiple imputation. In this thesis we study single imputation, in particular nearest neighbour methods, and we have developed a new method. Our method is based on dissimilarity measures and is nonparametric and handles categorical and continuous covariates without requiring any transformations. One drawback with this method was that it is relatively computer intensive, so we investigated data reduction methods.

For data reduction we developed a new method that uses propensity scores. Propensity score is used as it has properties that suggest that it would make a good method for matching the respondents and nonrespondents. We also looked at subset selection of the covariates using graphical modelling and principal component analysis. We found that the data reduction methods gave as good a result as when using all variables and there was considerable reduction in computation time especially with the propensity score method.

As the imputed values are not true values, estimating the variance of the parameter of interest using standard methods would underestimate the variance if no allowance is made for the extra uncertainty due to imputed data being used. We examined various existing methods of variance estimation, particularly the bootstrap method, because both nearest neighbour imputation and bootstrap are nonparametric. Also bootstrap is a unified method for estimating smooth as well as non-smooth parameters. Shao and Sitter (1996) proposed a bootstrap method, but for some extreme

situations this method has problems. We have modified the bootstrap method of Shao and Sitter to overcome this problem and simulations indicate that both methods give good results.

The conclusions from the study are that our new method of multivariate nearest neighbour is at least as good as regression based nearest neighbour and is often better. For large data sets, data reduction may be desirable and we recommend our propensity score method as it was observed to be the fastest among the subset selection methods as well as have some other advantages over the others. Imputing using any of the subsets methods we looked at appear to have similar results to imputing using all covariates. To compute the variance of the imputed data, we recommend the method proposed by Shao and Sitter or our modification of Shao and Sitter's method.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Nonresponse is a common phenomenon in large scale surveys and census. Survey organizations deal with the problem of nonresponse in different ways. In this thesis we propose some new methods for dealing with nonresponse based on one of India's large scale surveys, the National Family Health Survey. The author of this thesis, having been involved in the second of these surveys (NFHS-2), felt that the way that nonresponse was handled there could be improved and hence embarked on this project. In this chapter, a brief background of India's National Family Health Survey (NFHS) is outlined in section 1.1. In section 1.2, the sample design used in the NFHS is described. Computation of selection probabilities is given in section 1.3. Details of nonresponse in NFHS-2 are discussed in section 1.4. Research objectives and outline of the thesis are described in section 1.5.

1.1 The National Family Health Survey

India is a country with diverse cultures and languages. With a huge population, and large area, building policies on the basis of the decennial census is difficult due to

limited availability of information and the time gap between any two censuses. To formulate India's five year developmental plans, which require population projections, the Government conducts surveys such as NFHS. In the NFHS survey information on the social, economic, health and demographic status of the people living in India is collected. This information assists policy makers and programme administrators in planning and implementing strategies for improving population, health, and nutrition programmes. This survey was carried out by the Ministry of Health and Family Welfare, India, under the technical guidance of Macro International, with financial support of United States Agency for International Development. The NFHS survey has been conducted twice.

The first round of NFHS was conducted in the year 1992-93 and has become a major landmark in the history of demographic surveys, because of the availability of information at an individual and household level. The wide usage of the survey information by researchers, administrators, policy makers and planners created the need for the second round of this survey (NFHS-2), and this was conducted in the year 1998-99.

The general objectives of the NFHS-2 survey are;

- To provide state and national level information on fertility, the practice of family planning, infant and child mortality, and the utilization of health services provided to women aged 15-49 and children aged less than or equal to 3 years. In addition the survey provides indicators of the quality of health and family welfare services, reproductive health problems of women in the age group 15-49 and domestic violence.

- To provide information on the nutritional status of women aged 15-49 and children under 3. Height and weight measurements for younger children (less than or equal to 3 years) and women aged 15-49 are also collected. In addition

the rates of prevalence of anemia[1] is also provided for both women aged 15-49 and children less than 3 years of age.

The NFHS-2 sampled more than 90,000 eligible women from 26 states; representing more than 99 percent of India's population (IIPS, 2000). The sample is designed to provide urban and rural estimates of indicators such as fertility, mortality, nutrition status etc for most states. Regional estimates of desired indicators are given for the four big states, Bihar, Uttar Pradesh, Madhya Pradesh, and Rajasthan. Estimates for three metro cities- Kolkata, Chennai and Mumbai are also given.
The data from one of the states, namely Uttar Pradesh (U.P), is used in this thesis. Hence, the details of NFHS-2 for this state alone are presented. Hereafter NFHS-2 refers to the NFHS-2 survey for the state, U.P.

1.1.1 Data Collection in Uttar Pradesh

Uttar Pradesh (U.P) is located in the northern part of India, with Lucknow as its capital. The state has more than one-sixth of the total Indian population and one tenth of the land area of the country (IIPS, 2000). The state is divided into 19 administrative divisions and 63 districts. Geographically, U.P can be divided into five regions, namely Hill, Western, Central, Eastern and Bundelkhand [2](Census, 1991). Every region has distinct social, economic, and cultural characteristics. Though there are different local dialects, Hindi is the most commonly spoken language. For U.P, the objective of the survey is to provide information on indicators for regions, rural and urban areas separately and the state as a whole. In order to provide the information on these indicators, data was collected through personal interviews and health investigations.

[1]deficiency of red cells or deficiency of hemoglobin

[2]the region named after the Moghul King Bundelkhan

The NFHS-2 data collection process has six parts;

1. Select the primary sampling unit (PSU).

2. Select a secondary stage unit (SSU) in all the Urban PSUs and for few big Rural PSUs. For others the SSU is the PSU

3. Select households from the SSU.

4. Conduct household interview.

5. Conduct individual interview for any ever married women in the age group 15-49, termed as eligible women (EW) in the household.

6. Conduct health investigations for all interviewed eligible women and those of her child/children less than 3 years of age (hereafter child).

In the individual interview, eligible women were asked questions related to social and economic status, general health, nutrition, family welfare, reproductive health, and child and infant mortality. In addition, eligible women and younger children had their height and weight and hemoglobin levels measured. The hemoglobin measurement was performed with a Hemocue machine (Figure 1.1).
Selection of samples in this survey was made using a multistage stratified design. The details of NFHS-2 sample design for the state of U.P is described in the following sections.

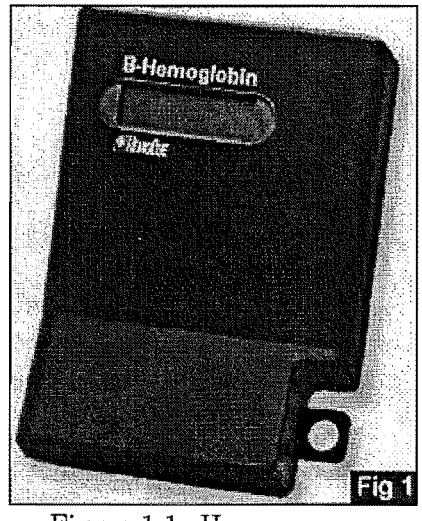

Figure 1.1: Hemocue

Source: www.hemocue.com

1.2 Design and Sample selection

1.2.1 Sample Size and Sampling Units

The targeted sample size is 10,000 completed interviews out of 15000 eligible women in U.P. This sample of 10,000 is to be taken from a total of 333 PSUs;

- The primary sampling units were villages for rural areas and wards[3] for urban areas.

- The SSUs were census enumeration blocks (CEBs)[4] in urban areas. In rural areas, only villages with more than 500 households required secondary stage selections.

[3]Wards are the administrative and electoral divisions in urban areas (Census, 1991).

[4]A CEB as defined by the Registrar General of India, is a block that comprises of 150-200 residential households (Census, 1991).

- The households were selected as the last stage. As every women reported eligible in the household was interviewed, no further selection was made.

The sample is designed to provide estimates of the indicators for the state as a whole, rural and urban areas, and for five major regions of the state (Hill, Eastern, Western, Central and Bundelkhand).

1.2.2 Sample Frame

The 1991 census list of wards was used as the frame for selecting the PSUs in urban areas. For selecting the PSUs in rural areas, villages with less than five residential households are deleted from the 1991 census list. Later some adjustments, such as linking the small villages, were made to create PSUs. Small villages with 5-49 households were linked to one or more adjoining villages to form a rural PSU with a minimum of 50 households. This linking was made by considering the administrative boundaries and the distances from the main village to which the small village was being linked.

1.2.3 Sample Selection

1.2.3.1 Sample Selection in Rural Areas:

Sample selection in rural areas was mostly done in two stages but for certain PSUs in three stages. For the first stage selection:

1. The frame was first stratified into five geographic regions (Hill, Central, Western, Eastern, and Bundelkhand) and the 63 districts assigned to them (Appendix-A).

2. The two biggest strata, Western and Eastern (S2 and S4 in Fig 1.2) were further subdivided into three smaller strata each.

3. In each geographic stratum, further stratification is done according to,

   - the number of residential households in a village;
   - the percentage of the population of scheduled caste or tribes [5]; and
   - the percentage of males engaged in nonagricultural activities.

4. The villages in this list thus obtained after stratification, were ordered according to the level of female literacy.

5. Finally from this list, PSUs were selected systematically with probability proportional to the population size of the village, obtained from the 1991 census.

Some villages which have more than 500 households, were split into three or more SSU. Two SSUs were selected using probability proportional to size (PPS) sampling. This selection of SSU was carried out at the house listing stage. The household sample was then taken in all the selected PSUs or SSUs as applicable.

1.2.3.2 Sample Selection in Urban Areas:

The sample selection in urban areas was performed in three stages,

1. Selection of PSUs (Wards)

2. Selection of SSUs (CEBs) from PSUs

3. Selection of households in SSU.

[5]The caste groups and tribes that the government of India officially recognizes as socially and economically backward and in need for special protection from injustice and exploitation (IIPS, 2000)

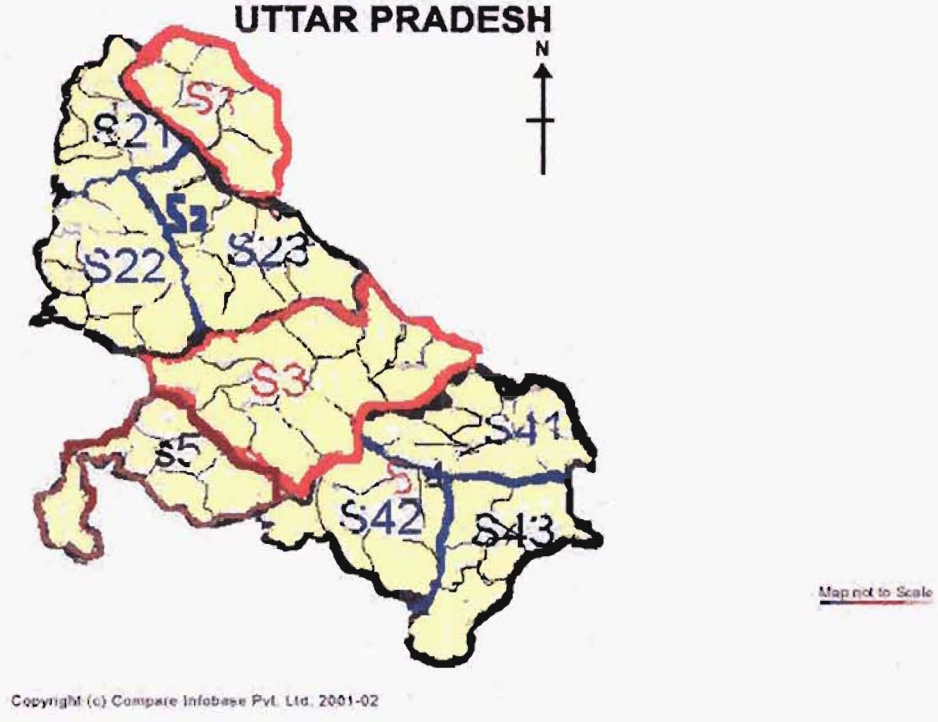

Figure 1.2: Stratification in Uttar Pradesh

Source: www.mapsofindia.com

For the first stage selection, the 1991 census list of urban wards was used as the sampling frame. For the sample selection in urban areas of the five regions the wards were arranged according to districts and within district by the levels of female literacy. From the list a sample of wards were selected systematically with probability proportional to the population size. A CEB was randomly selected from the sampled ward. Finally the households were selected within the sampled CEB.

1.2.4 House Listing and Household Selection

In order to select the houses in rural and urban areas the house listing operation was carried out in the selected village/CEB. In all the selected villages/CEB a team of two persons had to prepare a map which showed the exact boundaries of the village or CEB and identified the location of structures in the village/CEB. Here a structure is defined as a place which is intended for living by humans or cattle, or used as a warehouse, a place of worship or any other purpose such as schools and hospitals. This list of all the structures served as a frame for household selection. From this list of structures all the dwellings[6] are identified and renumbered after omitting the nonresidential structures. From this numbered list of dwellings, on average, a sample of 30 households were selected systematically in each village/CEB.

No replacement of the nonrespondent households was made. However if any PSU was inaccessible, a replacement was done by a PSU with similar characteristics, but this was rare. From the selected household all women identified as eligible were interviewed.

[6]A dwelling is defined as a place where a group of persons or single person live, eat and sleep together (IIPS, 2000)

1.3 Computation of Selection Probabilities

Computation of probabilities is done separately for Rural and for Urban domains. Generally the sampling fraction f is decided on for each domain and an n is determined from the domain by

$$f = \frac{n}{N}$$

where N is the projected population of eligible women as on September 1998 and n is the number of women to be interviewed in the domain. The number b of households to be sampled is also determined prior to the survey (and was 30 for both the urban and rural domain).

First we describe the computation for rural areas. A number $a = \frac{n}{b}$ is computed where a is the number of PSUs to be selected. b is the fixed number of households to be selected in the PSU or SSU depending on whether the area is non-segmented or segmented.

Having computed a, the probability of selecting a j^{th} PSU is calculated by

$$p_j^{(1)} = \frac{a * N_j}{\sum_j N_j} \tag{1.1}$$

where N_j is the population size of the j^{th} specific PSU within the rural area and $\sum_j N_j$ is the total population in the rural areas

From b we find that, for the usual case of 2 stage selection, the probability of selecting a household in the rural areas is:

$$p_j^{(2)} = \frac{f}{p_j^{(1)}} \tag{1.2}$$

For those rural PSUs which require a third stage of selection, the second stage selection probability is;

$$p_{j,\ell}^{(2)} = \frac{s * N_\ell}{N_j} \tag{1.3}$$

where s is the number of SSUs to be selected from a sampled PSU (in U.P s=2) and N_ℓ is the population size of a specified SSU ℓ.

The probability of selecting a household in such a SSU is computed as

$$p_{j,\ell}^{(3)} = \frac{f}{p_j^{(1)} * p_{j,\ell}^{(2)}} \tag{1.4}$$

Thus the overall probability of selecting a woman in rural PSUs that are not segmented is

$$f = (1.1) \times (1.2)$$

For rural PSUs that are segmented into SSUs the overall probability of selecting a woman in rural areas is

$$f = (1.1) \times (1.3) \times (1.4)$$

Computation of probabilities in urban areas is similar to rural areas with three stage selection. The only difference is in (1.3) s=1.

1.4 Survey Implementation and Nonresponse in NFHS-2

1.4.1 Implementation:

After the PSU or SSU where applicable was chosen the field work for NFHS-2 commenced. The data for NFHS-2 was collected through personal interviews. For these personal interviews three sets of questionnaires were used to collect information on households, individuals, and villages. All these interviews were conducted by a team. The team comprised a supervisor, an editor, a group of 4 female interviewers and a health investigator.

The duties of the supervisor are: to conduct the village interview, where the information on villages were collected from the head of the village, assist the female interviewers in identifying the selected households and look after the logistics of the team.

The duties of the editor are: to verify the consistency of the information, to guide the interviewers when they have difficulties during the data collection (for example, interpreting the questions without paraphrasing them. Convince the respondent to participate in the survey), and to reduce the loss of information (nonresponse, bad responses, loosing questionnaires) in the questionnaires.

The duties of the four interviewers are: to collect information on the household members and conduct the individual interviews.

The duties of the health investigator are: to take the measurement of height and weight of the interviewed eligible women and their children under 3 years of age, and also to measure their hemoglobin levels from a blood sample.

The author of this thesis, as a research officer for the eastern part of the state of Uttar Pradesh, was one of those responsible for the house listing training, individual training, and overall data quality monitoring. Every effort was made by the team to prevent non-response but in spite of this, nonresponse was unavoidable.

1.4.2 Nonresponse in NFHS-2:

The total number of PSUs/SSUs selected in the survey were 333. Of these 67 (20 percent) were urban and 266 (80 percent) were rural. From these 333 PSUs, 9626 households were selected. Interviews were completed in 90 percent of cases. The selected households were absent for an extended period in 5 percent of cases,

3 percent of the households were identified as not dwellings, and in 2 percent no household member or no competent respondent was at home at the time of the survey (IIPS, 2000). In the interviewed households 9292 women were identified as eligible for the individual interview.

The individual interviews had two types of nonresponse, unit nonresponse and item nonresponse. Unit nonresponse is when information for the individual interview is not given for all the survey questions. Item nonresponse is when information is missing for certain sections or questions in the questionnaire. The data on response show that 93 percent of the women identified as eligible from the household survey participated in the individual survey. According to the NFHS-2 state reports (IIPS, 2000), the breakdown of the 7 percent that did not participate is as follows:

- Four percent due to unavailability of the eligible women despite repeated household visits.

- 2 percent of the women refused to participate in the survey.

- Other reasons (1).

Item nonresponse in the health section was high compared to other sections of the questionnaire. In the health section the nonresponse for the hemoglobin measurement was on average 38% for the state. In contrast, for the rest of the variables that are not collected in the health section, the nonresponse was less than 5% on average for the state.

1.5 Research Objectives and Outline of Thesis

1.5.1 Research Objectives:

In the earlier section on nonresponse in NFHS-2, it was stated that NHFS-2 survey had two types of nonresponse; unit, and item nonresponse. As for most survey organizations, unit nonresponse in NFHS-2 was handled by adjusting the sample selection weights for nonresponse (Kish, 1965; IIPS, 2000). For item nonresponse weighting adjustment was also used. But the literature on item nonresponse suggests imputation is a better approach. Imputation is a technique by which the values of the nonrespondents are filled in by estimated values to make the data set complete. It can be seen that imputation is likely to make better use of what data the respondent has provided. Use and applicability of the existing methods of imputation to NFHS-2 data was critically examined. From this initial survey on imputation methods, it is felt that existing methods were not ideal for imputation of the Health Section of the NFHS-2 data. Hence for this thesis we aim to:

- Develop suitable methods for adjusting the item nonresponse in NFHS-2.
- Develop variance estimation methods under the proposed imputation method.

1.5.2 Outline of Thesis

This thesis is divided into seven chapters. Chapter 2 provides the details on the preventive measures taken to avoid nonresponse in NFHS-2 and an overview of the methods for handling nonresponse. Chapter 3 presents a new method for imputing the nonresponse in NFHS-2. Chapter 4 discusses data reduction methods that can make imputation more efficient. We also propose a new method for data reduction. Chapter 5 presents the results obtained from the application of the methods devel-

oped in chapters 3 and 4 and discusses them. Chapter 6 investigates the new variance estimation method particulary as applied to the new methodology introduced in the thesis. A summary and conclusions are presented in the last chapter.

Chapter 2

Nonresponse in Survey Data

In chapter 1, details of the National Family Health Survey (NFHS-2) design were presented along with a brief introduction to the types of nonresponse and percentage of nonresponse in the survey. Throughout this thesis we concentrate on item nonresponse, and adjustments to deal with item nonresponse. As defined in chapter 1, item nonresponse refers to the situation where the information is missing in certain sections or questions of the questionnaire. Particular attention is on item nonresponse in health related aspects (hemoglobin measurement) of NFHS. In this chapter, more details on item nonresponse in NFHS-2 are presented along with the type of response mechanism and an overview of the methods for adjusting the nonresponse in items. The measures adopted to reduce the item nonresponse and the reasons for nonresponse in NFHS-2 are presented in section 2.1. In Section 2.2, the classification of response mechanisms and the available methods for testing for a type of response mechanism are listed. A literature review on the available methods for adjusting item nonresponse is documented in section 2.3. A summary is presented in section 2.4.

2.1 Item Nonresponse in NFHS-2

"*Prevention is better than cure.*" The best strategy for dealing with any problem is to keep it from becoming too large (Lessler and Kalsbeek, 1992). Like other surveys NFHS-2 took preventive measures to avoid significant amounts of nonresponse. The main ones are given below.

2.1.1 Preventive Measures

At the time of questionnaire pretesting it was observed that nonresponse in health section of NFHS-2 was mainly due to reluctance of respondents to participate or equipment breakdown. With this in mind, the following measures were taken:

- Read out to the selected respondent the procedure for measuring the height and weight and of conducting the hemoglobin test;
- Demonstrate the process of measuring height and weight and taking blood samples with the help of health personnel such as auxiliary nurse midwife (ANM), or the health worker of the village;
- Instruct the field teams to have additional batteries and power outlets as the hemocue machines (Fig.1.1) worked both on batteries as well as on electricity.
- Instruct the health investigators to take double the number of cuvette's[1] as there were interviewing women.
- Make arrangements for a revisit the following day if the selected respondent was not free for health investigation. The maximum number of visits that were allowed are three, on various days and at various times.

[1]A small gadget (Fig.2.1) used for collecting a drop of blood

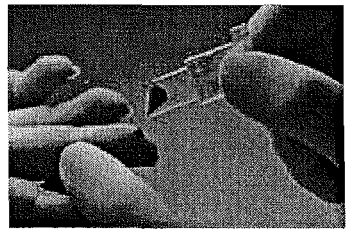

Figure 2.1: Cuvette

Source: www.hemocue.com

Despite the best efforts to prevent item nonresponse through the above preventive measures, nonresponse in items still appeared in some of the survey data. The nonresponse for the hemoglobin measurement on an average for rural and urban areas was 38 percent. As illustrated in figure 2.2, this item nonresponse can be due to missing by design, missing by logic, missing by routing or missing by refusal. The definition of these terms are:

- *Missing by design*: In every survey we take a sample from a population, thus leaving out other units and creating nonresponse. Such nonresponse is termed as missing by design. When this applies only to part of the survey such as the Health Section of NFHS-2, it results in item non-response.

- *Missing by logic*: The survey questionnaires are designed taking all possible types of respondents into consideration. If questions are not applicable to a particular respondent, they are supposed to follow a correct skip pattern and leave out certain questions. Such nonresponse is missing by logic.

- *Missing by routing*: Some times a person may be eligible for an interview, but follow a wrong skip pattern when answering the questions in the survey, thus

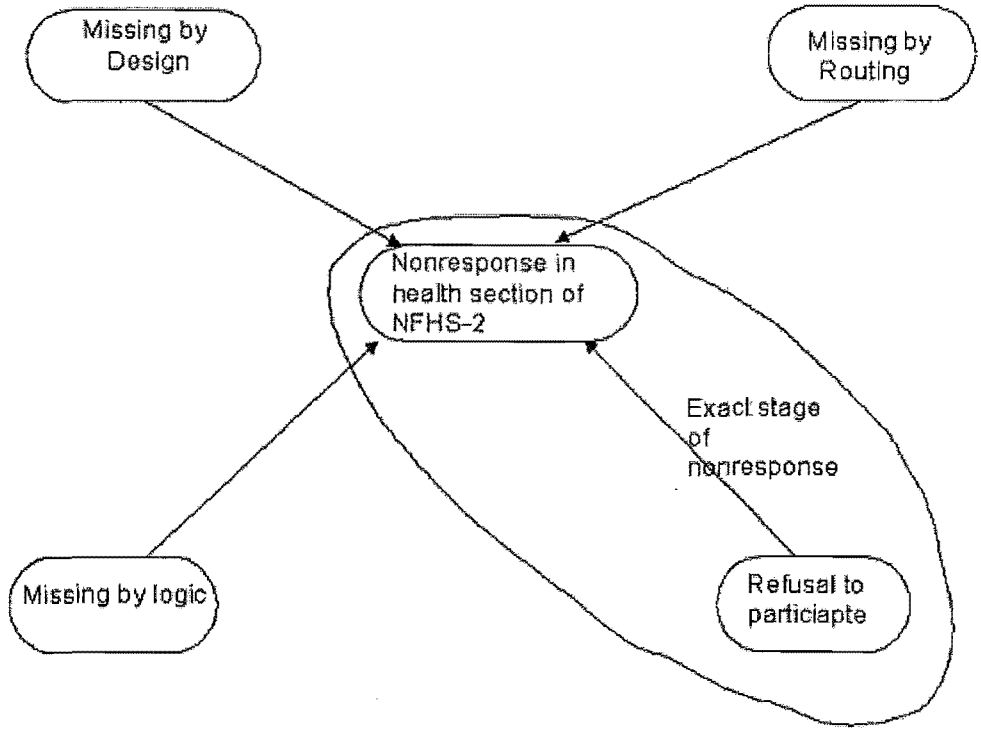

Figure 2.2: Stages at which nonresponse may occur

ending up answering wrong questions. This creates nonresponse in questions related their eligibility. Such nonresponse is missing by routing.

- *Refusal to respond*: When the selected respondent does not answer certain questions, perhaps due to sensitivity of questions, such as for example questions related to domestic violence. Such a nonresponse is called refusal to respond.

In NFHS-2 all the eligible women identified were interviewed for the health section, so there are none missing by design. Moreover, no skip pattern is observed in the questionnaire to examine the eligible women for her health measurements, and hence none can be missing by logic or accident. Hence we can conclude that the nonresponse in NFHS-2 is due to refusals.

2.1.2 Reasons for Refusals

Generally, the main reasons for nonresponse in health surveys are the respondent refusing to participate in the examination, for example because of the fear of the method of examination or the failure of machines (Lessler and Kalsbeek, 1992; Korn and Gourbard, 1999). Listed below are some of the reasons the author noticed for refusal to participate in health section of NFHS-2:

- Women afraid that their husbands would be angry at them being tested without the husband's consent.
- Women working on the farms refusing to get tested for fear that it would prevent them from working.
- Muslim women refusing to participate as the survey was being carried out

during the Ramadan[2] period.

- Women having a false notion that they are being tested for HIV/AIDS virus, and therefore refusing to participate.

Similar observations were made by Ferber (1966) who studied the effects of participant characteristics on item nonresponse. In his study he found that age, sex, occupation and eduction are strong correlates of item nonresponse. These suggest that the refusal to participate in the health section is influenced by factors like husbands approval, religious causes, sex, age, education and occupation.
The reasons given above generally influence nonresponse in the dependent variable Y. In addition the nonresponse may be a result of (Rubin, 1976):

- random reasons such as the interviewer not asking a question by mistake and thus independent of Y
- a currently pregnant woman not wanting to participate so that the nonresponse is correlated to Y
- or the nonresponse in Y variable is due to itself (for example, if women with low hemoglobin were not measured for their hemoglobin measurements)

This leads to the following classification of the response mechanism.

2.2 Classification of Response Mechanism

According to Rubin (1976), if the response in a variable Y is

- influenced by no known factors it is termed as missing completely at random (MCAR),

[2]Muslim festival of fasting

- influenced by any of the variables other than Y, then it is missing at random (MAR),
- related in some way to the study variable Y itself then it is not missing at random (NMAR).

To define these more precisely, let

$\mathcal{Z} = (\mathbf{X}, Y)$: be a data matrix with dimension equal to $n \times (v+1)$.

$\mathbf{X}$: be a set of covariates, dimension $n \times v$

Y: be the dependent variable whose dimension $n \times 1$. In this thesis we treat this as the variable with nonresponse.

$\mathcal{R}$: be the response indicator $= 1$ if Y observed and 0 if Y not observed.

ψ: be an unknown set of parameters of the response model.

The missing-data mechanism is denoted by $f(\mathcal{R}|\mathcal{Z}, \psi)$, the conditional distribution of $\mathcal{R}$ given $\mathcal{Z}$ and ψ. For our purpose we assume that $\mathbf{X}$ is completely observed and Y, the dependent variable, has nonresponse. There are three ways to define the response mechanism:

- Missing Completely At Random (MCAR): The missing data mechanism is termed missing completely at random if the distribution of $\mathcal{R}$ does not depend on the values of the data $\mathcal{Z}$, that is,

$$f(\mathcal{R}|\mathcal{Z}, \psi) = f(\mathcal{R}|\psi) \ \forall \ \mathcal{Z}, \ \psi. \tag{2.1}$$

  In this case nonresponse mechanism is not related to either covariates $\mathbf{X}$ or the dependent variable Y.

- Missing At Random (MAR): The missing data mechanism is called missing at random if the distribution of $\mathcal{R}$ depends on the covariates $\mathbf{X}$ That is

$$f(\mathcal{R}|\mathcal{Z}, \psi) = f(\mathcal{R}|Y, \mathbf{X}, \psi) = f(\mathcal{R}|\mathbf{X}, \psi) \ \forall \ Y, \ \psi \tag{2.2}$$

  This assumption is less restrictive than MCAR in that missingness depends only on the completely observed components of $\mathcal{Z}$ but not the dependent variable which has not been fully observed.

- Not Missing At Random (NMAR): The missing data mechanism is called not missing at random if it is not MCAR or MAR.

Examples of these three mechanisms are given in Little and Rubin (2002, 18-19). Understanding the distinctions between the three models is important, because they help one handle the missing data appropriately, and thus produce approximately unbiased estimates (Little and Rubin, 2002).

The values of the missing data are not known except by undertaking an extensive follow-up surveys which may be costly or impracticable. Thus generally, we cannot compare the observed values to the missing values to see the nonresponse model. Hence mechanisms like MAR and NMAR are impossible to test (Allison, 2001). However it is possible to test if we have MCAR. To test whether the missing data is MCAR or not Little (1988) has given a test. This test is available in many statistical software that have missing data methods (SPSS, VISTA, Mplus, SPlus).

2.2.1 Little's MCAR Test

Usually in a survey like NFHS-2, there may be many covariates and many dependent variables denoted in this section by $\mathbf{Y}$. If we assume that the covariates are fully observed, and only the dependent variables have missingness, we notice that the

data matrix $\mathbf{Y}$ could have holes in various places, thus creating different patterns of missingness for the variables $\mathbf{Y}$. Fig. 2.3 below illustrates a possible pattern of missingness, the ones filled in gray color are the observed data and ones in the black color are not fully observed. A pattern J is a set of variables and cases so that within J there is no missingness. We choose the patterns consecutively and at each stage ensuring that the maximum number of possible variables are included. For example, in Fig. 2.3 we have five different patterns. One way of assessing whether

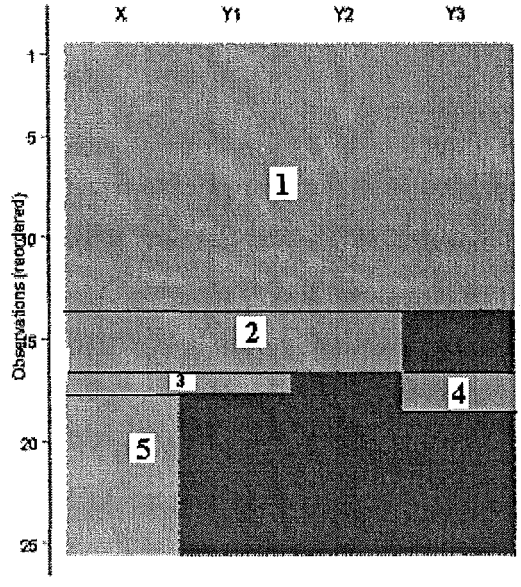

Figure 2.3: Nonresponse patterns

the missing data is MCAR or not is to compare the means of recorded values of each variable in each pattern.

Little's MCAR test is about testing the equality of means between the patterns. In

this test, for each pattern the pattern means ($\bar{Z}_{obs,j}$) are calculated first and then an expectation maximization (EM) algorithm (see Little and Rubin, 2002) is used to find the population mean $\mu_{obs,j}$ and covariance matrix $\Sigma_{obs,j}$. The pattern means and the EM estimates are used in the computation of d_0^2 where

$$d_0^2 = \sum_{j=1}^{\mathcal{J}} r_j(\bar{Z}_{obs,j} - \mu^*_{obs,j})\Sigma^{-1}_{obs,j}(\bar{Z}_{obs,j} - \mu^*_{obs,j})^T$$

where

- $\mathcal{J}$= total number of patterns,
- r_j= number of respondents in pattern j,
- *obs* refers to the observed cases

This d_0^2 follows a χ^2 distribution with the degrees of freedom equal to the difference between the number of means available across the data patterns and the number of variables i.e $v - J$. This test shows whether the means over patterns are the same, thus indicating whether or not the missing data is MCAR. If d_0^2 is significant then the data is MCAR.

2.2.2 Little's MCAR Test for NFHS-2

For the purpose of dealing with the missing data in NFHS-2, it is important to test whether it is MCAR. This test was performed using the statistical package VISTA, developed by Valero and Young (2000). For applying MCAR test on NFHS-2 data we used all the variables listed in Appendix-B. In NFHS-2 data there were 28 patterns of completely observed cases, we used all the patterns to compute Little' MCAR test. From these 28 patterns the pattern means were computed and the population estimates (means and covariances) are obtained using EM algorithm.

VISTA uses the change in the matrix of parameters as a criteria for the convergence of EM algorithm. The results of Little's MCAR test show that the test statistic was 11217.90. With the degrees of freedom equal to 1046 the p-value equals 0.0, thus showing that the missing data is not missing completely at random. Even though this test confirms that the missing data mechanism is not MCAR it does not offer a direct evidence on the validity of MAR compared to NMAR. In situations like this it is common to assume MAR mechanism. This allows one to choose an appropriate process to handle the missing data and produce approximately unbiased survey estimates. Having assumed that the data is MAR we now look at methods for handling data with nonresponse.

2.3 Methods for Handling Data with Nonresponse

Despite our best efforts to minimize nonresponse in NFHS-2, item nonresponse for the variable hemoglobin was as high as 38% in both rural and urban areas. Thus to perform analysis on hemoglobin requires us to find ways to deal with this nonresponse. The most commonly used methods for dealing with nonresponse are;

- Weighting adjustment;
- Case analysis; and
- Imputation.

2.3.1 Weighting Adjustment

The idea is to weight the data from the respondents such that they make up for the effect of nonrespondents. This is done by computing the sample adjusted weights for

each variable with nonresponse. Two common methods of item level nonresponse weighting adjustment are

1. weighting class adjustment and

2. propensity weighting

(Oh and Schuren, 1983; Lessler and Kalsbeek, 1992; Little and Rubin, 2002).

2.3.1.1 Weighting class adjustment

Under this approach the whole sample is divided into H classes on the basis of homogeneity of observed variables; each class with its set of respondents and non-respondents. In each class h, the response probabilities for each case i, (ϕ_h) are computed by dividing the respondents (r_h) with the sample size (n_h) in each class (i.e. $\phi_h = r_h/n_h$). Once these probabilities are estimated, they are multiplied with the sample selection probabilities (π_{hi}) of each case in each class. The nonresponse adjustment weight (w) for each case is then computed as

$$w_{hi} = \frac{r(\pi_{hi}\phi_{hi})^{-1}}{\sum_{i=1}^{r}(\pi_{hi}\phi_{hi})^{-1}} \; \forall i \in \mathrm{R}$$

(Little and Rubin, 2002, page 47). This adjustment procedure was used in NFHS-2 survey results (IIPS, 1998).

2.3.1.2 Propensity Weighting

Weighting class estimators can be applied when the set of observed covariates $\mathbf{X}$ is small. However when the set of covariates is large, construction of weighting classes becomes difficult. If all the information on $\mathbf{X}$ is available for all the respondents and nonrespondents, then Little (1986) advocates forming adjustment classes using propensity scores and then applying the above weighting procedure. An alternative

is to weight the i^{th} respondent by the inverse of the estimated propensity (Cassel, Sarndal and Wretman, 1983).

2.3.1.3 Issues with Weighting

Under both these procedures each variable with missing data has a corresponding set of adjustment weights to be used in analysis. This idea of using separate case weights for each variable with nonrespondents highlights one of the main limitations of this method.

1. Computing weights in this manner will be time consuming since the adjustment method chosen for each variable must be applied separately; and

2. An analyst doing multivariate analysis will have difficulties in choosing which case weights to use (Lessler and Kalsbeek, 1992).

Although the weighting adjustment methods have some good statistical properties (e.g. preserving correlations and joint distributions), they still require relatively careful analysis for each item that has missing data. With item nonresponse the missingness is not a planned missingness as in missing by design. This can result in excessive variability even though the estimates may be unbiased (Rubin, 1996). For item nonresponse the weights are computed from the observed data and hence are themselves subject to sampling uncertainty. The influence of ignoring this source of variability when computing the standard errors is very unclear (Little and Rubin, 2002). These problems imply that complete data methods may not lead to valid inferences.

2.3.2 Case Analysis

A common approach of handling missing data is to just use the complete cases for analysis.

2.3.2.1 Complete Case Analysis

This is also known as list wise deletion. In complete case analysis the cases with nonresponse in any of the variables are discarded. If the MCAR mechanism holds, the observed cases can be treated as a random sub-sample of the actual sample and data analysis procedures can be used for finding the statistics of interest such as means, totals and variances (Little and Schenker, 1995; Little and Rubin, 2002).

2.3.2.2 Available Case Analysis

Here there are two approaches used for estimating the statistic such as means, totals and variance (Little and Rubin, 2002).

- In case of estimating the variance of a single variable we use all the available cases for that variable.
- In case of estimating the variance for a pair of variables we use the number of cases that are available for both these variables.

This is also known as pairwise deletion.

2.3.2.3 Advantages and Disadvantages of Case Analysis

The advantages of these approaches are that:

- they are simple and standard complete data statistical analysis without any modification can be applied and

- since the statistic of interest is computed on a common set of complete cases they are comparable.

The disadvantages of these approaches are that:

- there is a potential loss of information in discarding the incomplete cases and this is a major disadvantage
- If the missing data mechanism is not MCAR, then the nonrespondents are not a random sub-sample of all the cases, hence there is a loss of precision in the estimate and bias.

In our study we have shown the missing data mechanism for NFHS-2 is not MCAR, therefore this simple method for handling missing data could provide misleading results. Therefore, we look at imputation as an alternative.

2.3.3 Imputation

Assume we have fully observed covariates $\mathbf{X}$ an $n \times v$ matrix. We also have a variable Y, an $n \times 1$ vector. Y has missing values, so can be divided into $Y = (Y_{obs}, Y_{mis})$; where *obs* are the respondents and *mis* are the nonrespondents. The covariates associated with the Y variable can similarly be divided $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$. Note that $\mathbf{X}_{mis}$ are not missing covariates but rather the known covariates corresponding to Y_{mis}.

The current practice in large scale surveys is to handle unit nonresponse (both $\mathbf{X}$ and Y are missing for the same case) by weighting and item nonresponse (Y alone is missing as described above) by imputation. The basic reason for imputation is to make the data set look complete, that is, with no holes in it. This presents a complete data set to the analysts. It should also reduce the bias due to nonresponse. Imputation aims to:

- find a replacement value for Y_{mis} that is as similar as possible to the true value of the missing case;
- Reduce the biases of estimators and preserve the relationships between the variables;
- Provide consistent results to various analysis (Kalton, 1981).

These imputation methods can be broadly classified into two categories:

1. Deterministic
2. Stochastic.

Deterministic and stochastic imputation as defined by Rubin (1996) and Bello (1994) are:

(Deterministic Imputation:). *Imputation methods that are repeatedly applied on the same data produce the same imputed values.*

(Stochastic Imputation:). *Imputation method that incorporates a random error and thus different imputed values may be produced when repeatedly applied to the same data set.*

The main advantages of a deterministic methods is that it produces a single data set. This is appealing to the data analyst as the data analyst will have no difficulties in using complete data methods on the imputed data which when repeatedly applied gives the same results.

A disadvantage is that the variance of the estimate obtained from such imputed data is underestimated. Stochastic methods attempt to overcome this problem by adding an error term to the imputed value. This addition gives variability in the

imputed values such that the variance due to imputation is not underestimated (Little and Rubin, 2002; Lohr, 1999; Govindarajulu, 1999; Allison, 2001; Ford, 1983). The disadvantage with stochastic imputation is that multiple data sets are generated so that it is hard to choose a data set on which further analysis can be done. Another disadvantage is that there may be problems with storage.

2.3.3.1 Cell Mean Imputation

Similar to the weighting class adjustment, the respondents are divided into H classes based on known variables. The mean of the values of respondents in class h, $\bar{y}_{hr}$, is used to impute all the nonrespondents in that class h (Kalton and Kaspryzk, 1982). When the number of classes is equal to one then this method is equivalent to the simple mean imputation method. Cell mean imputation may be a feasible choice where analysis is limited to simple point estimates without variances. Although cell mean imputation is easy to perform and reasonably effective in reducing the bias in point estimates such as population means and totals, it has some disadvantages. Firstly, cell mean imputation distorts the distribution of the Y variable in that, the value for all nonrespondents in a particular cell is always the sample mean of the responding cases in the sample. This creates a spike in the distribution of Y (Lohr, 1999). Secondly, as there is little variation among the sample members in an imputation cell, the variances of the point estimates may be underestimated (Lessler and Kalsbeek, 1992).

2.3.3.2 Cold Deck Methods

The word deck is from the days when punch cards were used to store computer data. The deck was hot if the cards were taken from the current survey. By contrast if the imputed values are not taken from the same survey, it is called cold deck. Imputed

values are usually obtained from a previous study or from other information such as historical data (Little and Rubin, 2002; Lohr, 1999). Little theory exists for this method. Cold deck methods do not guarantee to eliminate selection bias (Lohr, 1999). Cold deck methods are deterministic methods. One example of cold deck imputation is;

- Exact matching: In exact matching the imputed values for the nonrespondents are taken from the records of the same unit, but from another source (e.g. health, administrative, or tax records) (Lessler and Kalsbeek, 1992). In this situation unique identifying information, such as a social security number, drivers license number, is used to match a nonrespondent. Studies by Schieber (1978), Cox and Bonham (1983), Platek and Gray (1983), of this method find it a good method when the external source of information for nonrespondents is available. However there may be instances where the information in external data source may not be entirely consistent with the information one is trying to collect (e.g. a persons income for tax purpose may not be equivalent to their actual income). Many surveys are collecting information for the first time (e.g. hemoglobin measurement in NFHS-2). In such situations this method cannot be applied.

2.3.3.3 Hot Deck Methods

Hot deck imputation is a very generic term used to describe a family of widely used imputation methods. A hot deck method is one in which each missing value is replaced by the value from a similar case that responded in the same survey (Lohr, 1999, Lessler and Kalsbeek, 1992, Little and Rubin, 2002). The choice of the imputed value is made within homogenous subsets of the sample. A detailed discussion on hot deck imputation method for large scale surveys is in Fellegi and

Holt (1976), Sande (1983) and Rizvi (1983). Choosing a complete case (donor) for missing value (recipient) under the hot deck method can be done in several ways differing in how the donor is chosen. Common methods are sequential, random and nearest neighbour.

- **Sequential hot deck Imputation:** This is a deterministic imputation method. In this, records are ordered using a covariate that is highly correlated with the variable that has nonresponse. The idea behind ordering is to create a data set where consecutive records in each imputation class or cell are as similar as possible with respect to Y variable. From the ordered set the value of the previous card with response is used as a donor for the recipient. One problem with using the value on the previous card is that often nonrespondents tend to occur in groups, so one person may be a donor multiple times.

- **Random hot deck Imputation:** A donor for the recipient is randomly selected from the respondents of the same class. In those cases when there is missingness in more than one variable to preserve the multivariate relationships, a single donor is often used to impute all the missing items for a nonrespondent (Lohr, 1999). This method may be made stochastic if the selection process is random every time an analyst imputes the values for nonrespondents. Random hot deck overcomes the problem of sequential imputation if the random draws assume simple random sampling without replacement. But this methods does not make use of covariate information, hence may not considerably reduce the nonresponse bias, especially when the data not MCAR as we have shown the case for NFHS-2.

  To make use of the covariate information there are some hot deck methods such as nearest neighbour imputation.

- **Nearest Neighbour Imputation:** This method is also known as distance function matching. All possible donors are identified in terms of a quantifiable measure of distance to the nonrespondent. The donor is that respondent with the least distance from the recipient. Distance is measured as a function of the covariates (Little and Rubin, 2002; Lohr, 1999; Lessler and Kalsbeek, 1992; Godfrey *et al*, 2002)). Nearest neighbour imputation requires logical choices in measuring *nearness*. Sande (1983) in her paper outlines some possible ways of defining the *nearness* between the respondents and nonrespondents. Recent work by Chen and Shao (2000) presents the theoretical properties of the nearest neighbour methods, where they show that nearest neighbour imputation methods are better than other hot deck imputation and regression imputation. Some of the most commonly used distance measures are:

  - Caliper distance: Consider the simple case where a single covariate (X) is used to identify a donor. Here *nearness* between the i^{th} and j^{th}, where $i \in obs$, $j \in mis$, sample members is:

    $$d_{ij} = |X_i - X_j|$$

    When the distribution of covariates are skewed then transforming the variables to make the distribution symmetric is recommended (Rubin, 1987). The nearest neighbour obtained for the missing case j is the case k for which $d_{kj} = \min_{1 \leq i \leq r}(d_{ij})$. For this method it is necessary that X is continuous.

    This method is used by many survey organizations such as Statistics Canada, U.S. Bureau of the Census, U.S. Bureau of Transportation (Rancourt et-al, 1994) and Statistics New Zealand.

  - Mahalanobis distance: When covariates are multivariate and continuous,

then one can use Mahalanobis distance. The Mahalanobis distance is one of the oldest distance measures and it is used in many multivariate analysis. Vacek and Ashikaga (1980) and Little and Smith (1983) in their work on edit and imputation used this distance for identifying the outliers prior to regression imputation. To compute the distance between any two sample members the Mahalanobis distance is,

$$\mathcal{D} = (\mathbf{X}_i - \mathbf{X}_j)'\Sigma^{-1}(\mathbf{X}_i - \mathbf{X}_j)$$

where Σ is the estimated covariance matrix for the set of covariates and $\mathbf{X}_i$ and $\mathbf{X}_j$ are respectively, the vectors of covariates for the i^{th} and j^{th} samples. Mahalanobis distance works well when all the covariates are continuous. With categorical variables this method tends to lose its efficiency (Rosenbaum and Rubin, 1983).

2.3.3.4 Regression Imputation

In regression imputation the missing values are obtained as follows:

1. Fit a multiple regression model using all the observed cases.

$$E(Y_{obs}) = \mathbf{X}'_{obs}\beta \tag{2.3}$$

   where $\beta = (\beta_0, \beta_1.....\beta_v)$ and $\mathbf{X}'_{obs} = (1, \mathbf{X}_{obs})$. β_0 is the intercept term.

2. Obtain the estimates $\hat{\beta}$ of parameters β, using equation (2.3);

3. Use these estimates of the parameters $\hat{\beta}$ from step 2 and the covariates $\mathbf{X}_{mis}$ to predict the missing values i.e.

$$\hat{Y}_{mis} = \mathbf{X}'_{mis}\hat{\beta} \tag{2.4}$$

(The above regression method can be regarded as stochastic imputation, when an error term is added to eq.2.4. i.e

$$\hat{Y}_{mis} = \mathbf{X}'_{mis}\hat{\beta} + \epsilon \tag{2.5}$$

The ϵ term can be defined in different ways. For example $\epsilon \sim N(0, \sigma^2)$ where σ^2 is the variance of the observed values of Y or the regression residual sum of squares (Kalton, 1981). If any of the covariates are categorical variables then they are transformed into dummy variables and used in the regression imputation (Little and Rubin, 2002). A special case of regression imputation when the intercept is zero and there is one variable is

$$\hat{\beta} = \frac{\Sigma_{i \in obs} Y_i}{\Sigma_{i \in obs} X_i}$$

this is ratio imputation.

2.3.3.5 Ratio Imputation

Like nearest neighbour imputation and regression imputation, ratio imputation also uses the covariates for imputing the missing data. As before, the covariate X is assumed to be completely observed (i.e. n cases) and is correlated to Y which is observed only for r cases. The imputed values for the $m(= n - r)$ nonrespondents are;

$$Y_j = \frac{\Sigma_{i \in obs} Y_i}{\Sigma_{i \in obs} X_i} * X_j$$

Single X is obtained from the set of covariates ($\mathbf{X}$), by looking at its correlations with Y. A covariate which is highly correlated with Y is used for imputing the missing values. A highly correlated covariate can yield accurate imputations (NCES report, 1999; Rancourt *et al*, 1994).

Even though ratio imputation can provide accurate imputation it has some draw-

backs. One drawback is that it is not known which variable to use for cases where there are several correlated variables (Rancourt *et al* 1994; Chen and Shao, 2000).

2.3.3.6 Semi-parametric or Hybrid methods:

Apart from a wide variety of individual imputation methods such as those above, there are some methods that are a mixture of both regression and hot deck methods. These methods are called hybrid or semi parametric methods. Little (1986) defined predictive mean matching method which is a combination of hot deck method and regression imputation. This method uses both categorical and continuous covariates (**X**). The method is as follows

- A regression model is fitted as in regression imputation (sec 2.3.3.4)
- The estimates of the regression coefficients from the previous step are used to predict the values of Y for both respondents and nonrespondents
- Make use of the predicted values of Y, Y^p (corresponding to the respondents and nonrespondents) to find the nearest neighbour. That is

  $$d_{ij} = |Y_i^P - Y_j^P|$$

  The nearest neighbour obtained for the missing case j is the case k for which $d_{kj} = \min_{1 \leq i \leq r}(d_{ij})$.
- Once the nearest neighbour is found then the observed Y value of the nearest neighbour is used as the imputed value for the recipient.

As would be expected, this method works well if the model is a good model of the data.

2.3.3.7 Summary of Single Imputation methods

All the imputation methods we have described so far are single imputation (except (2.5)) methods. That is, only one complete set of imputed data is obtained. There are advantages and disadvantages with these imputation methods. The advantage with some of the hot deck methods, cold deck methods and regression methods are that they use the covariate information collected to impute the missing values. If the imputation model is a good representation of the nonresponse model then it will reduce the nonresponse bias compared to complete case or weighting adjustments. They are also relatively easy to implement and provide a single clean data matrix whose values are mostly the observed values in the survey rather than predictive mean values. Another advantage with single imputation methods is that the standard complete data methods can be applied on the imputed data. If the imputation model is assumed to be correct, it provides good parameters estimates. However the single imputation methods do not account for the uncertainty due to imputation (Little and Rubin, 2002). This leads to a systematic underestimation of the standard errors computed from the imputed data (Rubin, 1987) and thus the statistical significance of the analysis may be wrongly estimated. To resolve this problem Rubin (1977) proposed multiple imputation.

2.3.4 Multiple Imputation (MI)

The basic idea of multiple imputation (Rubin, 1977, 1987) is:

1. Impute missing values using an appropriate stochastic imputation model;

2. Repeat this M times, producing M "complete" data sets;

3. Find the estimates of interest (e.g. means, total) from each data set;

4. Average these M values of the estimates from the MI samples to produce a single point estimate;

5. Compute the between-imputation variance and within-imputation variance using the formulas given in Rubin (1987). Combine these two variances to get the total variance of the parameter of interest.

The main advantages of multiple imputation are that:

- Good estimates of standard errors are obtained due to repeated imputations;
- Depending on the method used for creating multiple data sets, multiple imputation can be performed for any kind of missing data patterns without any specialized software (Rubin, 1987).
- Incorporating an appropriate random error into the imputation process makes it possible to obtain approximately unbiased estimate of the parameter of interest.

More details on multiple imputations are given in Rubin (1987). Multiple imputations are done under the assumption of an ignorable missing data mechanism. The missing data mechanism is an ignorable mechanism when the data are missing at random and the data generating parameters and the response generating parameters are distinct (Little and Rubin, 2002, pp 119-120);
Under the ignorable missing data mechanism, Rubin and Schenker (1986) presented multiple imputation procedures for discrete and for continuous variables. Schafer (1997) has suggested some mixed models for data that has both continuous as well as categorical variables.
For discrete variables, the methods suggested by Rubin and Schenker (1986) are Bayesian Bootstrap and Approximate Bayesian Bootstrap . For more details refer

Rubin (1987) or Govindarajulu (1999). In order to impute the continuous data Rubin and Schenker (1986) proposed two methods; fully normal imputation, and imputation adjusted for uncertainty in the mean and variance.

A commonly used multiple imputation method that makes use of covariates is data augmentation. Data augmentation is a form of Gibbs sampling where Gibbs sampling is a special case of Markov chain Monte Carlo methods (MCMC). The MCMC method is a technique for creating pseudorandom draws from probability distributions. More details on data augmentation and other related multiple imputations can be obtained in Tanner and Wong (1987).

Even though multiple imputation procedures have some advantage over single imputation one of the complications for multiple imputations is according to Allison (2001) and Schafer(1997):

Using additional variables in the imputation process: Suppose the imputer uses a subset of covariates ($\mathbf{X}$) for imputation, whereas the analyst uses all the covariates in his later analysis then, the inferences that use the standard methods may not be valid.

2.4 Conclusions

It has been shown using Little's MCAR test that the missing data mechanism for NFHS-2 is not MCAR. This shows that imputing the missing data is desirable. The basic approach for dealing with an ignorable nonresponse is to adjust the nonresponse using the covariate information. This helps reduce the bias in the estimates. This assumption that the nonresponse is ignorable allows one to develop techniques that can easily be programmed to account for the observable differences. Assuming MAR we have described several methods of imputation which have their own advan-

tages and disadvantages. In this thesis we choose to use single imputation methods rather than multiple imputation for the following reasons.

- The data set from NFHS-2 will be used by many people from various fields of research and varying statistical skills. It is, hard to form an imputation model that represents the needs of analysts.

- Variance due to single imputation can now be captured using the recent techniques developed by Rancourt *et al* (1994), Chen and Shao (2001), Rao and Shao (1992).

- The ability to predict the missing values close to the true values may be more adversely affected by a poor imputation model than by the use of single value imputation methods (Landerman *et al*, 1997).

Of the single imputations discussed in the above sections we intend to use nearest neighbour (NN) methods to the semi parametric methods. We choose to use nearest neighbour because it has been proved by Chen and Shao (2000) that the biases are less compared to other hot deck imputation methods. In addition NN imputation method uses all relevant covariate information in the data when finding the nearest neighbour. Details on how the nearest neighbour is constructed is explained in the following chapter and we describe a new a nearest neighbour method that we developed for dealing with situations where there is a mixture of categorical and continuous variables.

Chapter 3

A New Imputation Method

3.1 Introduction

In chapter 2 we discussed some preventive measures taken by NFHS-2 to avoid nonresponse. Despite these, item nonresponse was unavoidable. To adjust for this item nonresponse, weighting adjustment method was used in NFHS-2 (see 2.3.1). But weighting adjustment methods reduce the data to complete cases analysis and this is only appropriate if the data is MCAR. We have shown that the NFHS-2 data is not MCAR, hence using these methods may give biased estimates and so we look at imputation methods. Imputation allows the use of standard complete data methods and, assuming the imputation model is correct, provides good parameter estimates. We choose single imputation over multiple imputation because of operational difficulties in maintaining, supplying and analyzing multiple complete data sets, especially when the surveys are large (Rao, and Shao, 1992; Yansaneh *et al*, 1998).

Of the single imputation methods, we use the stochastic imputation methods because it makes it possible to get approximately unbiased estimates of the parame-

ters. Among the stochastic single imputation methods we use nearest neighbour imputation. This is preferred over other methods because it makes full use of the covariates and is non parametric. Studies by Rancourt (1999) and Chen and Shao (2000) showed that the biases for nearest neighbour imputation are less compared to other hot deck methods. In addition, nearest neighbour imputation is a common method used by various organizations such as Statistics Canada, Statistics New Zealand, and US Census Bureau.

Section 3.2 of this chapter describes some candidate distance measures that could be used to find a nearest neighbour. Details on nearest neighbour imputation method that is commonly used in literature is described in section 3.3. Section 3.4 describes a new nearest neighbour imputation method which we have developed. Extensions of this method to situations where there are multiple variables with nonresponse is outlined in section 3.5. Section 3.6 discusses the possible modification of this method to weight different variables in the distance function to allow for their known importance in determining a donor. Section 3.7 discusses the details of different MAR response mechanisms that are used in this study. Lastly section 3.8 summarizes this chapter.

3.2 Distance Measures for Nearest Neighbour

Nearest neighbour imputation finds a donor (respondent) for a recipient (nonrespondent) using covariates. To achieve this a suitable distance measure is required to define nearest neighbour (that is, the closest in characteristics to the donor). Generally, a suitable distance measure d_{ij} between cases i and j is defined (see 2.3.3.3) and the nearest neighbour obtained for the missing case j is the case k where k is such that $d_{kj} = \min_{1 \leq i \leq r}(d_{ij})$.

When the covariates are multivariate and continuous then one can use measures such as Mahalanobis distance to find nearest neighbour. With categorical variables this method tends to lose its efficiency (Rosenbaum and Rubin, 1983).

However, when the data has both continuous as well as categorical variables, there is no standard approach to measure distance. In missing data analysis Little (1986) has suggested the use of a hybrid method (see 2.3.3.6). An example of this method is regression based nearest neighbour (RBNN) (Laaksonen 2000). The advantage with hybrid methods is that they handle mixed (categorical and continuous) type of variables, and defining distance to find the donor record is straight forward (see below). But the categorical variables need to be transformed when used in regression, and this transformation can lead to a loss in information (Kaufman and Rousseuw, 1990). Moreover, the quality of the nearest neighbour is dependent on the predictive power of the regression.

To address these limitations we describe a new method (see 3.4) which allows for different types of variables. First we review the RBNN methods which we use as a benchmark for evaluating our new method.

3.3 Regression Based Nearest Neighbour (RBNN)

Following the same notation we used in chapter 2 section 2.3.3.4, the steps in RBNN imputation are;

1. Fit a model using the respondents with complete information

$$E(Y_{obs}) = \mathbf{X}'_{obs}\,\beta \tag{3.1}$$

2. Use the estimate of β, $\hat{\beta}$, to predict Y

$$Y^P = \mathbf{X}'\,\hat{\beta} + \epsilon \tag{3.2}$$

3. The nearest neighbour is computed using the Y_{mis}^{P} and Y_{obs}^{P} rather than all the covariates.

4. Use the distance measure

$$d_{ij} = |Y_i^P - Y_j^P| \ \forall i \in obs, j \in mis$$

to obtain the nearest neighbour for the missing case j as the case k where $k : d_{kj} = \min_{1 \leq i < r}(d_{ij})$.

5. Use the Y_{obs} corresponding to the nearest neighbour obtained from step 4 as the imputed value for Y_{mis}

If ϵ in eq.(3.2) is assumed to be zero we have deterministic regression. However we retain this error term so we can avoid the problem of a spike in the distribution of Y at a particular value. This results in the standard errors of the imputed variable not being biased (Landerman *et al*, 1997; Shao and Wang 2002). If the data sets are small and there is high proportion of nonresponse, it may not be advisable to impute, as in these circumstances the choice of error term can make a difference to the estimation of variance of the parameter of interest obtained from the final data (Allison, 2001). In addition by adding an error term it can be shown that the difference between the sampling variance calculated on data after imputation, and the unknown variance of a sampling consisting entirely of actual observations is approximately zero on the average (Sarndal, 1990). A common practice is to assume that the error term $\epsilon \sim N(0, \sigma^2)$. For generating the error term we require σ^2. According to Kalton and Kaspryzk (1982) σ^2 can be obtained in several ways. Here in this thesis we follow Laaksonen (2000) where σ^2 is the mean square error of the regression eq.(3.1).

3.3.1 Advantages and Problems with RBNN Methods

Some of the advantages of RBNN method for imputation are that:

- It does not tend to underestimate variance as it is stochastic, rather than deterministic.
- It is likely to perform better than either simple regression or hot deck imputation alone (Cochran and Rubin, 1973; Laaksonen 2000),
- The imputed value is a value observed in the survey rather than a predicted value.
- For categorical variables, we do not get decimal values as imputed values, as we could get in a simple regression imputation (Laaksonen, 2000; Grau *et al*, 1999; Landerman *et al*, 1997).

Some of the disadvantages are that:

- When there are various types of covariates, finding a good regression model can be difficult.
- When the sample size is small and the number of categorical variables are many, then there is a need to create many dummy variables. This may cause a singularity when computing $\hat{\beta}$.
- Failure of multivariate normality assumption may lead to heavy tailed distribution of the imputed values (Schafer, 1997). This in turn can lead to overestimation of variance.
- With a decrease in the predictive power of the imputation model the possibility of finding an appropriate substitute value decreases. By predictive power we

do not only mean that R^2 = say 0.9 is necessary, but that the model must have variables that can explain the variation in the dependent variable as well as the response variation.

To overcome the problems of explicit models and not to loose the information by transforming categorical variables, we develop a new method of nearest neighbour imputation which is nonparametric. This method we call multivariate nearest neighbour imputation (MVNN).

3.4 Multivariate Nearest Neighbour Imputation

3.4.1 General Idea

When data is multivariate and the variables are not all continuous, there have been no distance measures used for missing data analysis. For standard multivariate analysis, distance measures have been used for discriminant and cluster analysis (see e.g. Krzanowski, 1983, 1987) Kaufman and Rousseuw (1990) have used a metric they called *a dissimilarity* for their work on cluster analysis. We adapt that here for missing data analysis.

- Let c and $\mathfrak{m}$ be cases whose covariates are observed.
- Compute the distances (d_{ij}) as given in (3.4) between the cases using the distance method appropriate for the type of variable as described below.
- subsequently take a selected sum of all the distances for all variables as in equation (3.3).
- This process is repeated for all the cases in the data. This will lead to a $n \times n$ matrix of dissimilarities $[D(c, \mathfrak{m})]$

- For the missing case $\mathfrak{m}$, choose from all possible donors the case with the $D(c, \mathfrak{m}$ as the nearest neighbour and is to be used as the donor. If there are several donors with the same $min[D(c, \mathfrak{m})]$if a donor may be randomly selected from them.

3.4.2 Details of the Method

For the purpose of an exposition we assume that the missingness is in one variable though we extend this to multiple nonresponse later in the chapter. The dissimilarity between any complete case c and a missing case $\mathfrak{m}$ is defined as

$$D(c, \mathfrak{m}) = \frac{\sum_{j=1}^{v} \delta_{cm}^{j} d_{cm}^{j}}{\sum_{j=1}^{v} \delta_{cm}^{j}} \tag{3.3}$$

where the distance d_{cm}^{j} is

$$d_{cm}^{j} = \begin{cases} \left.\begin{array}{ll} 1 & \text{if } x_{cj} \neq x_{mj} \\ 0 & \text{otherwise} \end{array}\right\} & \text{for binary and nominal} \\ \frac{|x_{cj} - x_{mj}|}{\rho_j} & \text{interval and ordinal variables} \end{cases} \tag{3.4}$$

δ_{cm}^{j} is an indicator variable which takes 1 for all variables except where the j^{th} covariate is asymmetric and $x_{cj} = x_{mj} = 0$, in which cases $\delta_{cm}^{j} = 0$. An asymmetric binary variable is one where the outcomes are not of equal importance in terms of their predictive power. For comparison a symmetric binary variable, say where sex = 1 is male and 0 is female, the categories 1, and 0 are of equal importance for dissimilarity computation. However, when we consider the variable "termination of pregnancy" of the NFHS-2 survey, we have two categories 1 and 0, where 1 corresponds to "never terminated pregnancy" and 0 corresponds to "ever terminated pregnancy (etp)". Here "etp" could include terminations that have occurred recently, or in many years the past. Of these it is terminations that occurred most

recently that are likely to be more important than ones in the past for the purpose of finding a donor for hemoglobin nonresponse. However, this information cannot be obtained from the simple binary coding from NFHS-2. In this the more important outcome is coded as 1 and the less important as 0. That is 1 has more importance than 0. Hence in the case for terminated pregnancies outlined above a 1-1 match of an asymmetric binary variable is more important than a 0-0 pair in terms of choosing a donor. As the 0-0 pair does not provide useful information for matching, it should be disregarded in the computation (Kaufman and Rousseeuw, 1990, p.26).

We use ρ_j, the range of the j^{th} covariate, rather than standard deviation to normalize so that the interval and ordinal variables have a distance d^j_{cm} in [0,1], consistent with other variables. Ordinal and ratio scaled variables are treated as interval scaled variables.

With this approach, the dissimilarity computation takes care of all types of variables to define a suitable nearest neighbour for use as a donor record.

3.4.3 General Comments on the Dissimilarity Metric

The dissimilarity,

- is a nonnegative number.
- matrix is symmetric.
- will be zero for case to itself.
- computed for n cases would be $\binom{n}{2}$
- has all the properties of a metric except for triangle inequality, although this is not an essential property for imputation.

- of all continuous variables; if we redefine the distance d^2_{cm} as a squared distance and use σ_j^2, the variance instead of range ρ_j, then

$$d^2_{\text{cm}} = (x_i - x_j)'\sigma_j^{-2}(x_i - x_j)$$

and so is similar to Mahalanobis distance.

3.5 MVNN with Missingness in Several Variables

Large scale surveys collect information on various variables that would be of interest to a wide range of analysts. Many of these variables may have nonresponse, some of the nonresponse can be in the dependent variables and some can be in the independent variables which we call covariates. Here dependent refers to the variable of interest and independent refers to the variables that are treated as covariates in a particular study (for example in this thesis Hemoglobin is dependent variable and other variables are covariates). Hence it is necessary that an imputation method be able to impute all these variables. This helps preserve the joint distributions.
Extension of the MVNN method can be done in two ways;

- Assume monotone pattern (See Fig 3.2) of nonresponse.
- Assume a more general pattern of nonresponse

3.5.1 Monotone Pattern of Nonresponse

Arrange the variable according to the size of the complete cases, then sort the data according to completely observed cases, such that the ordering may possibly create a monotone pattern like that in the Fig(3.2). If the data has a monotone pattern it can be easy to handle. Apply the MVNN method directly without any modification

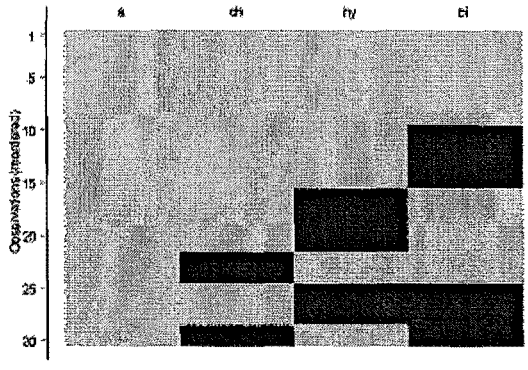

Figure 3.1: General nonresponse pattern

to complete variables and impute the nonresponse in the dependent variable with the least number of missing cases. Treat these imputed values as the true values and use this variable as a covariate for imputing the next variable with the least number of missing cases and so on for all the remaining dependent variables.

Generally in practice, the nonresponse pattern is rarely monotone, but we may be able to get close to a monotone pattern and assume it to be monotone. When data do not follow a monotone pattern or cannot be approximated to it as above then one can use the following method which make necessary adjustments to eq (3.3).

3.5.2 General Pattern of Nonresponse

In surveys it is hard to observe a monotone pattern of nonresponse all the time. This is because nonresponse does not limit only to a dependent variable, there may be more than one dependent variable with nonresponse or covariates may also have nonresponse. In such situations nonresponse is at irregular places in the data. This

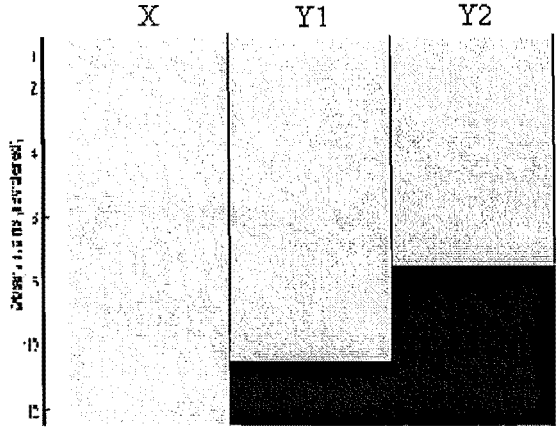

Figure 3.2: Monotone nonresponse pattern

creates patterns such as shown in Figure-3.1. To extend the MVNN method to handle more general patterns of nonresponse, we need an additional indicator I_{cm} that identifies the variable which has nonresponse and removes it from the computation of dissimilarity. Equation (3.3) is then modified to

$$D(c, m) = \frac{\sum_{j=1}^{v} I_{cm}^{j} \delta_{cm}^{j} d_{cm}^{j}}{\sum_{j=1}^{v} I_{cm}^{j} \delta_{cm}^{j}} \tag{3.5}$$

where

$$I_{cm}^{j} = \begin{cases} 1 & \text{if } x_{cj} \text{ observed} \\ 0 & \text{otherwise} \end{cases}$$

3.6 Weighting MVNN for Important Variables

In some survey we know that some variables are more important than others. In these cases we can modify our method to account for their importance by adding a weight to the distance function. Thus each variable j will have associated with

it a weight w_j and the distance function can be extended to include these case by redefining it as

$$D(c, \mathrm{m}) = \frac{\sum_{j=1}^{v} w_j I^j_{\mathrm{cm}} \delta^j_{\mathrm{cm}} d^j_{\mathrm{cm}}}{\sum_{j=1}^{v} I^j_{\mathrm{cm}} \delta^j_{\mathrm{cm}}} \tag{3.6}$$

3.7 Weighting MVNN for Complex Designs

MVNN can be further extended to accommodate complex survey designs by modifying (3.6) by adding the design weights π_i as follows

$$D(c, \mathrm{m}) = \frac{\sum_{j=1}^{v} \pi_i w_j I^j_{\mathrm{cm}} \delta^j_{\mathrm{cm}} d^j_{\mathrm{cm}}}{\sum_{j=1}^{v} I^j_{\mathrm{cm}} \delta^j_{\mathrm{cm}}} \tag{3.7}$$

3.8 Simulated Response Models

To study our proposed methodology we will use simulations to compare its performance with Laaksonen's (2000) recommended stochastic imputation methodology RBNN. For these simulations we create artificial nonresponse using several missing at random (MAR) response mechanisms described below. It is hard to measure the performance of an imputation method using the survey data at hand. This is because most of the values for the nonrespondents are not known and to conduct a follow-up survey of the nonrespondents to get their responses is not feasible. Instead simulated nonresponse experiments is usually performed. With simulations one has the advantage in knowing the true values of missing data and the knowledge on how the nonresponse is generated. Thus one can measure the performance of the imputation methodology under defined circumstances.

Our simulations are done using observed NFHS-2 data. To truly test the performance of imputation methods one needs to use several response models. This allows us to identify the situations when a method works well. In this study we

used four different MAR mechanisms, simple MAR, MAR linear, MAR convex and MAR concave. The MAR linear and MAR concave are used by Collins *et al* (2001). The MAR convex is defined here.

3.8.1 Missing at Random: Simple

As reported in chapter 2, missing at random (MAR) means the probability to respond depends on the covariates but not the dependent variable. In order to generate a simple MAR mechanism we used two covariates, religion and current pregnancy status as the covariates which model nonresponse. We set nonresponse in hemoglobin (HL) if

$$[((\text{religion} = \text{Muslim}) \cap (\text{pregnancy status} = \text{currently pregnant})) \cup (U < \vartheta)]$$

where U is an independent random variable uniformly distributed over [0,1] and ϑ is a constant that is used to modify the probability that a case has missing value.

3.8.2 Missing at Random: Linear

In MAR-linear, the probability of missingness is linearly related to a covariate. An example reported in Collins et-al (2001) would be a survey where individuals with higher values on the covariate income have higher probabilities of nonresponse to a question on the use of financial services. To achieve this we divided our covariate (X) into H classes and assigned response probabilities to each class in a linear pattern (see fig 3.3). We multiply these probabilities with the sample size in each class to get an estimate of the number of nonrespondents m in each class h. We then take a random sample size equal to m_h and insert nonresponse in HL corresponding to these m_h cases.

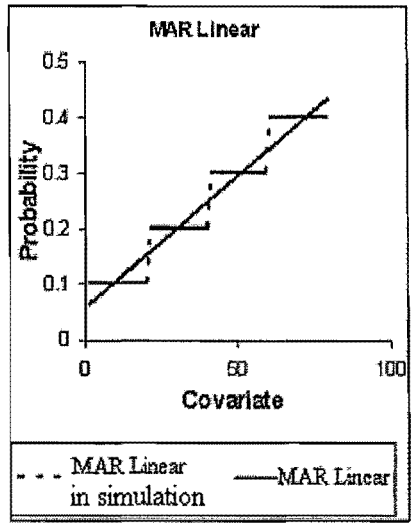

Figure 3.3: MAR linear mechanism

3.8.3 Missing at Random: Convex

To illustrate this, let us take the example of incomes and utilization of financial services used in previous section. Suppose instead of the linear relationship between nonresponse and income that we had previously, we have persons from low and high incomes more likely to not to respond to the question on utilization of financial services than the ones with the middle income. In MAR-convex, the probability of missingness is higher in the first (H/4) and last (H/4) classes and lower in the middle (H/2) classes. To achieve this a similar procedure as described in MAR linear was adopted. Instead of using probabilities in an increasing manner we used high probabilities in the first and last classes and smaller probabilities in the second and third classes (see fig 3.4)

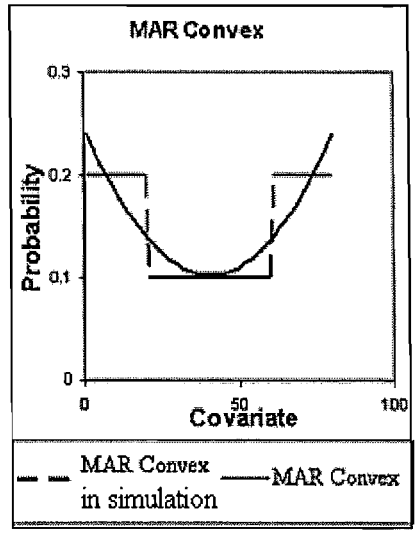

Figure 3.4: MAR convex mechanism

3.8.4 Missing at Random: Concave

In contrast to MAR convex, in MAR-concave, the probability of the missingness is high in the middle (H/2) classes but low at the end (H/4) classes. Here we used high probabilities of nonresponse in the middle classes and low probabilities in the top and bottom classes (see fig 3.5).

3.9 Accuracy

To create simulated data for imputation, nonresponse was generated in the NFHS-2 data using one of the four response mechanism outlined above. The data with simulated nonresponse was then imputed using our MVNN and the benchmark RBNN method. The performance of these methods was tested for their ability to impute the values close to the true values. The accuracy of an imputation method

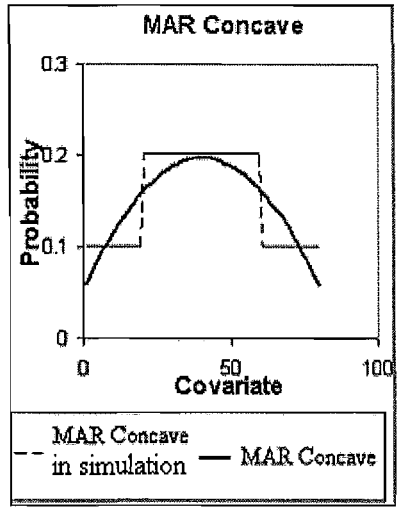

Figure 3.5: MAR concave mechanism

can be assessed in many ways, as there are number of ways in which the data could be analyzed. In this thesis we have chosen to:

1. Compare the error in the imputed values,

2. Find the RMSE of the imputed values,

3. Compare the error in the marginal distribution of the imputed data with the true distribution,

4. Find the RMSE due to imputation for the parameter estimates,

For the first and third comparisons we used the methods given in Manzari and Reale, (2002) and Leti (1983). The second and fourth comparisons do not appear to have been used elsewhere.

3.9.1 Individual Imputations

Accuracy of the individual imputations for numeric variables is measured by mean absolute deviation.

$$\Delta = \frac{\sum_{j=1}^{m} |y_j^* - y_j|}{m} \tag{3.8}$$

where y_j is the actual value taken from the sample data set before creating non-response and y_j^* is the imputed value. The smaller the value of Δ the better the performance of the imputation method in terms of estimating the nonresponse values close to true values. This is a good measure for a single imputation class, where an imputation class are homogenous groups constructed using observed covariates (for example, age, religion, etc). But when there are several imputation classes this is not a good measure (see next section 3.9.2). Hence we look at the mean square error of the imputed values.

3.9.2 Mean Square Error of the Imputed Values

The mean square errors of imputed values are computed as follows;

$$RMSEI = \sqrt{\frac{\sum_{j=1}^{m} (y_j^* - y_j)^2}{m-1}}$$

The advantage of using the RMSEI over Δ is that suppose there are H imputation classes and the imputation is done separately in each class so that there is no cross use of respondents across the classes. Suppose the mean absolute deviations (Δ_h) are computed for each class $h = 1, 2...H$ separately. Then generally $\frac{1}{H}\sum_{i=1}^{H} \Delta_h \neq \Delta$ (Youden, 1951). But this is not the case for the mean square error and so might be a preferred measure when using imputation classes.

This measure indicates the closeness between individual imputed values and true values but does not describe the preservation of distributions.

3.9.3 Imputed Marginal Distributions

The accuracy of imputed marginal distributions was studied using an index given by Leti (1983). This computes the difference between the relative distributions of imputed values and the actual values. For categorical variables the index can be applied directly. For numeric variables, first we categorize by dividing it into h classes before computing the index. The index is:

$$\Gamma = \left(\sum_{\ell=1}^{h} \frac{(|g(\ell) - f(\ell)|)}{2} \right) * 100$$

where $g(\ell)$ is the cumulative relative frequency of the ℓ^{th} category in the imputed data, $h(\ell)$ is the cumulative relative frequency in the actual data. The limits for Γ are (0-100). When there is no difference in the relative distribution between imputed and actual data then Γ is zero. When the Γ value is 100 then there is maximum differences in the relative distributions. While it does not measure very well the large relative differences in the proportions for a class, this measure gives a good overview of the imputed values.

3.9.4 Mean Square Error of Parameters with Imputations

In surveys, the final objective is to present the parameter estimates of the dependent variable. With nonresponse, there is an additional bias due to imputation which cannot be estimated because the true values of nonrespondents are not observed. If the imputation model is not good then this bias would further increase. In simulated experiments we have an opportunity to study the bias and then decide on the performance of the imputation method. In order to find the differences in parameter estimates (e.g. means, ratios, totals) for the actual and imputed data we

use mean squared errors given by

$$MSE = \frac{1}{M-1}\sum_{s=1}^{M}(f(y)_s^* - f(y))^2 \tag{3.9}$$

where M is the number of simulations and $f(y)$ can be any linear function of Y. The lower the MSE the more consistent is the imputation method in providing estimates close to true estimates.

3.10 Summary

In this chapter we proposed a new method of imputation where the nearest neighbour for imputation is obtained using dissimilarity measure rather than a distance measure. This new method, which we call multivariate nearest neighbour (MVNN) uses both categorical as well as the continuous variables which other imputation techniques either cannot handle or require making compromises. This method described is detailed for a situation where the nonresponse is in one variable but extension of this method for the data sets with nonresponse in more than one variable is also discussed. We also described the measures that we will use to asses its performance as compared to our benchmark (RBNN) method. Results of these comparisons are given in chapter 5 and they show that our method performs better than RBNN for simple MAR models. For other MAR models considered in this thesis it is as good as RBNN method. Overall it is nonparametric and hence avoids model misspecification.

Chapter 4

Data Reduction before Imputation

4.1 Introduction

A detailed description of some of the existing imputation methods and a newly proposed imputation method is presented in chapters 2 and 3. Some of the methods such as simple random hot deck and mean imputation methods do not make use of any covariates, whereas methods like regression and nearest neighbour imputation do. Use of covariate information is helpful in reducing the nonresponse bias especially when the missing data mechanism is MAR (Rubin, 1987; Chen and Shao, 2000). Rubin also recommends including all available covariate information to its fullest extent, this is because

- Using all covariates will increase the predictive power of the model used for imputation
- Even though some covariates are not significant in the model they may be of subject importance.
- In practice the response model is not known hence if all the covariates are used

there may be less of a chance of leaving out the covariates that are the cause of nonresponse.

However when the sample size is small, even a simple model will be over parameterized as the number of variables may become more than the number of cases (Song and Belie, 2004; Schafer, 1997). In a multivariate regression model with a general covariance matrix, efficiently estimating the covariance matrix becomes difficult especially when the number of covariates is large relative to the sample size. Moreover, when some of the variables are collinear, the inverse of the variance covariance matrix might not exist as the matrix may become singular indicating that the estimates of the regression parameters may be imprecise (Schafer, 1997). In such cases, analysis often proceeds by choosing an arbitrary subset of the variables and or parameters. Most surveys conducted for social science or economic research collect information on various aspects of the survey unit. This results in many variables in the data. For modelling purposes not all the information collected in the survey may be relevant with respect to a particular dependent variable, hence variable reduction may be a preferred choice to select relevant information.

Because the number of variables in a data file may be too large, the time and effort to find an efficient subset of highly correlated variables for each dependent variable may be too large. Therefore if a "predetermined set" of covariates is selected using subject knowledge and experience from pervious studies, then data reduction on this predetermined set may be easier. We now investigate methods of choosing an appropriate subset of covariates from a predetermined set of covariates. There are several techniques in multivariate analysis that can reduce high dimension data set to a low dimension one without disturbing the main statistical features of the data set. Selection of the predetermined set of variables on which data reduction will be performed is presented in section 4.2 along with a need for data reduction

on the predetermined set. An overview of commonly used data reduction methods for selecting variables before imputation is presented in section 4.3. Subset selection methods used for comparison are presented in sections 4.4 and 4.5. Section 4.6 presents a new method of data reduction. Conclusions from this chapter are in section 4.7.

4.2 Selection of "Predetermined" Covariates

As described in chapter 1, information in NFHS-2 is collected at village, household, and individual levels. In this study we are interested in the individual information. This information was collected from all ever married women who were in the age group 15-49, termed as eligible women (EW). In the EW interview, the information was collected on aspects of the women's background information (e.g. age, rural/urban, etc), history of her fertility status, history of family planning methods, knowledge on AIDS (Acquired Immune Deficiency Syndrome), reproductive health, general health (tuberculosis, asthma, etc), quality of health care, domestic violence, and history of child immunization. Apart from these details, information on anemic status and body mass index was collected from hemoglobin and height and weight measurements. Since we are interested in imputing the nonresponse in hemoglobin level measurement, the selection of "predetermined set" of covariates was done using the subject knowledge from medical studies on anemia by Shilpa Sapre (2001), Sood *et al* (1975), Agrawal *et al* (1999) Kanani (1994) UNEP Report, 2002; Massawe, 2002; NFHS-2, 2000 and [http://www.reutershealth.com/wellconnected/doc57.html]. The following two tables present a description of how we arrived at the relevant sections for finding a match for imputing the hemoglobin variable. Table-4.1 presents the overall divi-

sion of the women's questionnaire. In addition the women's questionnaire contained detailed child information. Child information like abortion of a child and recent delivery of the child may help arrive at good imputation value. However other detailed child information collected in the survey were left out for two reasons:

- If a women has for example, three children it is hard to decide which child's information is to be included in the analysis.

- In the child questionnaire information was collected on immunization and knowledge on diarrhoea. This information may not affect the conclusions that we make on women's hemoglobin level.

Table-4.2 presents the details on the women's information, which was collected from a personal interview and medical examination. On the basis of the medical studies presented earlier the third column in table-4.2 indicates whether the information is considered relevant or not for imputing the missing data in hemoglobin. Thus the final data set used in this thesis for data reduction simulations has the covariate information from the sections which have "Yes" in table 4.2

Table 4.1: Division of questionnaire

Women's information
Background
Hemoglobin
Height and weight

Table 4.2: Relevant information chosen from the 2000 variables in the women's section

Women's information	Number of variables	relevant information (Yes/No)
General information	11	Yes
General Health information	3	Yes
Reproductive health information	2	Yes
Family planning	-	No
Fertility information	3	Yes
Nutrition information	7	Yes
Domestic Violence	-	No
Altitutde	1	Yes
Quality of Care	-	No
Knowledge on Aids	-	No
Health	2	Yes

4.2.1 Need for the Study

A desirable feature for a particular imputation method to be considered as a good imputation method by a survey organization, is that the method be easy to program and not computationally intensive. The new MVNN method we developed is, as stated in Murthy *et al* (2003) and in this thesis, relatively computationally intensive and so we will investigate the performance of the imputation methods by taking a subset of variables. In addition large number of covariates makes it difficult to find matched pairs with similar covariates. Hence variable reduction on the predetermined set of covariates may be useful to avoid difficulties in imputations. The motivation behind taking a subset of variables comes from the studies by Collins *et al* (2001) and Sixten and Sarndal (2002). In the study by Sixten and Sarndal they comment that even if technically feasible, it is not necessary to use all available covariate information. Subject knowledge and statistical association may be used for selecting the subsets. In order to identify and select a subset of covariates they provided the following guidelines:

i) The selected subset explains the variation in the response probabilities.

ii) The selected subset explains the variation of the dependent variable.

If i) is satisfied, then the MAR condition given in eq.(2.2) holds for the subset of covariates and the imputation using the subset of covariates reduces the nonresponse bias. If i) not satisfied, the nonresponse in the subset becomes MCAR and the estimates obtained using the subset will be biased especially when the nonresponse mechanism of the data is not MCAR. If ii)is satisfied, then the predictive power of the imputation model increases and thus may impute missing values close to true values. Hence for a subset to be representative of the full set of covariates both conditions i) and ii) should be satisfied. Collins *et al* (2001) observed that if the

subset of covariates can explain the variation in the response probabilities then the use of that subset or the full set of covariates did not make much differences in the estimates of the parameter of interest. Using these guidelines in this study we further select a subset of variables from the predetermined set to be used for imputing the dependent variable.

4.3 Data Reduction Methods

Use of data reduction for selecting a subset of covariates before imputation dates back at least to the studies of Dear (1959). Here subset of covariates refers to two situations

1. Reducing the set of v covariates to w covariates where $w < v$.

2. Form v' linear combinations which encapsulate the information of the v variables and then take a subset w from these new set of covariates v'.

In this section we present an overview on some of the most commonly used methods for data reduction by various survey organizations and researchers.

Chi-Square Automatic Interaction Detection method: The chi-square automatic interaction detection (CHAID) covariate data reduction method was used in Statistics New Zealand prior to imputing Māori descent [1] variable (Westbrooke and Jones, 2000) . The CHAID segmentation process, using the chi-square statistics, first divides the data into groups based on categories of the most significant covariate of the dependent variable and then splits each of these groups into smaller subgroups based on other predictor variables. The CHAID

[1] Māori are the native residents of New Zealand.

process may also merge the categories of a variable that were not found significantly different. This splitting and merging process continues till no more statistically significant predictors are found. In general CHAID is used for predicting the outcome of a dependent categorical variable on the basis of a set of categorical covariates (du Toit *et al*, 1986).

Ridge prior method: In some studies there are larger number of variables than the number of observations, or strong relationships between the variables may exist, creating collinearity. In this situation making inferences about the population parameter under a model based approach is difficult because the variance-covariance matrix ($\mathbf{S}$) may be singular. To overcome this difficulty Schafer (1997) suggests a method called "Ridge prior". In this method instead of looking at the data reduction of variables, he suggest reduction of the parameters in the multivariate model. This is achieved by transforming the original covariance matrix to a diagonal matrix, whose elements are the diagonal elements of $\mathbf{S}$.

Factor analysis Methods: Studies by Lee *et al* (2003) show the use of factor analysis for data reduction before imputation. In their study they select the subset of covariates using factor scores. These variables thus selected were then used for creating the adjustment cells (see 2.3.1.2). In another study by Song and Belin (2004) the use of factor analysis for data reduction is recommended. In their study they used the factor scores instead of variables and performed imputation.

From this overview we observe that survey organizations and individual studies use data reduction methods for reducing high dimension data set to low dimension data before imputation. With this in mind we investigate three data reduction methods.

The details of these three methods are presented in the following sections.

4.4 Principal Component Analysis

Principal component analysis (PCA) is one of the well known techniques of data reduction in multivariate statistical analysis. When using PCA the subset selection is done as described in situation 2 of section 4.3. Jolliffe (1972, 1973) was one of the first to investigate subset selection of variables by making use of principal components in the context of multivariate analysis. Later McCabe (1984) proposed some new techniques called selection of principal variables. The method of Jolliffe was proved to be inefficient by Cadima and Jolliffe (2001), who propose alternate solutions to this problem.

In imputation studies, Dear (1959) and Hu and Salvucci (2001) used principal component analysis for data reduction. In their approach they reduce the data to a principal component that explains most of the variation in data and use it for finding the nearest neighbour. Even though the dimension of the data can be reduced from v variables to 2 or 3 principal components, one may still need to interpret the results on the original variables (McCabe, 1984; Jolliffe, 1972, 1973; Jackson, 1991). Because principal components are a combination of variables it is difficult for the analyst to work out how the imputation model could affect the results of any particular analysis they do.

For nearest neighbour imputation we prefer a subset of covariates, not principal components. This is because interpretations of the final results will be easier. In addition it makes it appealing for the analyst to relate the variables used in imputation to the models they use for their specific analysis. Therefore in this study, we make use of two approaches proposed in Cadmia and Jolliffe (2001) for data reduction.

These are generalized coefficient of determination (GCD) and the subset of variables as predictors (SVP). GCD is essentially the average of the squared canonical correlations between each principal component and the set of selected variables (Ramsay and Silverman 1997). SVP is the weighted average of the square multiple correlation between each principal component and the set of selected variables, where the weights are simply the eigen values of the covariance or correlation matrix ($\mathbf{S}$). For more details refer Cadima and Jolliffe (2001).

The use of correlation matrix is recommended to obtain the principal components. This is to avoid the problems of sensitivity of measurements used for each element of the covariate data. Here we use both correlation as well as covariance matrix to obtain subset of covariates as a means of comparison. One might think that the subsets must be the same because it might seem that the PC's for correlations matrix could be obtained fairly easy from the corresponding covariance matrix since the correlation matrix can be derived from covariance matrix. However this is not the case. The eigen values and eigen vectors of the correlation matrix have no simple relation with those corresponding to the covariance matrix (Jolliffe, 2002). Hence we get a different subsets of variables.

Even though PCA is the most commonly used data reduction method it has some advantages and limitations. One of these is selecting the number of covariates to be retained in the subset. Graphical modelling avoids this problem and is one of the methods we investigate in this thesis.

4.5 Graphical Modelling (GM)

Graphical models are those probability models for multivariate random observations whose independence structure is characterized by a graph.

Use of graphical methods for subset selection was proposed by Falguerolles and Jmel (1993) in the context of multivariate analysis. We have applied this to data reduction for imputation. In this section we give an outline of graphical modelling. For more details see Whittaker (1990) Lauritzen and Wermuth (1989).

. *Graph: A structure with a finite set of edges ($\mathcal{E}$) connecting a finite set of variables called vertices or nodes ($\mathcal{V}$), is called as graph (G) (Edwards, 2000).*

Graphs are of two types: directed and undirected. In an undirected graph, the edges are unordered pairs; that is, each edge merely connects two vertices. In directed graphs the edges are ordered pairs; that is for each edge one vertex follows from another (Edwards, 2000). Initially graph is drawn with all the vertices and the edges connected. The edges for the variables that are conditionally independent will be omitted (Whittaker, 1990).

For example, suppose we have a set of three covariate X_1, X_2, X_3 and a dependent variable Y. Then complete graph would be similar to that in figure 4.1 but with a edge from X_3 to Y. In the graph below we assume that $Y \perp\!\!\!\perp X_3 | X_1$, hence the edge between Y and X_3 is omitted and if that is the only conditional independence relationship the graph is in its final form as shown below. In section 5.5.2 we discuss the way conditional independence is determined.

To draw a graph we use partial correlations computed for all the variables. For our work we use undirected graphs but it would be interesting to look at more details of the relationship between the variables Zio *et al* (2004); however this is felt to be of secondary importance and we do not look at that in this thesis. In selecting a subset of variables using graphical modelling we look for conditional independence relation of the form $Y \perp\!\!\!\perp X_1 | (X_2, X_3, ...X_v)$.

Graphical methods have the following advantages:

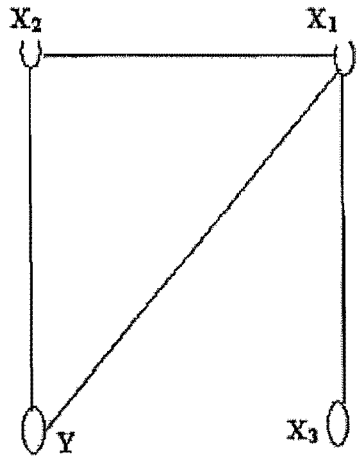

Figure 4.1: Undirected graph

- The graph is a pictorial representation, which is a very informative medium of communication.

- Graphical modelling can use both categorical variables and continuous variables.

- Graphs display the associational and causal dependencies between the variables in the model.

Even though graphical modelling is very informative medium in explaining how many edges are retained in the subset still is not free from problems. One of this is that it is not clear whether condition ii) given in the guidelines of Sixten and Sarndal is satisfied or not for the subset selected.

We look next at propensity as a method of data reduction; it does not have the above problem.

4.6 Propensity Matching

The theory of propensity score was initially given by Rosenbaum and Rubin (1983) (hereafter RR in this thesis) in the context of observational studies where they discussed propensity matching as a device for data reduction and finding homogenous matched sets, when $\mathbf{X}$ has many covariates. The propensity score(π) is defined as the conditional probability of responding to an item given the covariates. That is

$$\pi(\mathbf{X}) = pr(\mathcal{R} = 1|\mathbf{X})$$

(Rubin, 1985)has shown that propensity score is a sufficient summary of the covariates. In other words $\pi(\mathbf{X})$ is sufficient for θ if the conditional distribution of $\mathbf{X}|\pi(\mathbf{X})$ is independent of θ where θ is a set of unknown parameters to be estimated from $\mathbf{X}$. Use of propensity score for survey nonresponse was proposed by Little (1986) where he used propensity score for forming weighting cells (section 2.3.1.2). The following are the properties of propensity scores given by Rosenbaum and Rubin (1983) in context of observational studies.

1. $\pi(\mathbf{x})$ is a balancing score.

2. Nonrespondents and respondents with a matched set of propensity score tend to have the same distribution.

   Since Little's (1986) paper, no further application of propensity scores to missing values appears to be in the literature. Later in Little and Rubin (2002) they showed that:

3. Response indicator and covariates are conditionally independent given a propensity score (Little and Rubin, 2002); i.e. for knowledge of $\mathcal{R}$, information on X is irrelevant once $\pi(\mathbf{X})$ is given.

The term balancing score in above is defined as follows;

. *Balancing score: A balancing score, $b(\mathbf{X})$, is a function of the observed covariates* $\mathbf{X}$ *such that the conditional distribution of* X *given* $b(\mathbf{X})$ *is the same for nonrespondents and respondents Rosenbaum (2002); (RR)*

The following proposition show how the distribution of covariates are equal for the responding and nonresponding cases given the propensity score. The situation considered here is a special case of the proposition given in (RR) and Rosenbaum (2002). In Rosenbaum it was proved for the situation where propensity score was used for stratification, whereas we use propensity score for matching. We prove this proposition to show that the nearest neighbour obtained for the nonrespondent will have similar distributions of covariates to the respondent within the neighbourhood of $\hat{\pi}(\mathbf{X})$. Let $\mathbf{x}$ be a particular value of $\mathbf{X}$ and π determined by the response.

Prop 1. *Balancing property of propensity scores: If* $\pi(\mathbf{x}) = \pi$, *then*

$$pr[\mathbf{X} = \mathbf{x}|\pi(\mathbf{X}) = \pi, \mathcal{R} = 1] = pr[\mathbf{X} = \mathbf{x}|\pi(\mathbf{X}) = \pi, \mathcal{R} = 0] = pr[\mathbf{X} = \mathbf{x}|\pi(\mathbf{X}) = \pi] \tag{4.1}$$

Proof. If $\pi(\mathbf{x}) = \pi$, then by Bayes theorem

$$\begin{aligned} pr(\mathbf{X} = \mathbf{x}|\pi(\mathbf{X}) = \pi, \mathcal{R} = 1) &= \frac{pr(\mathbf{X}=\mathbf{x}, \pi(X)=\pi, \mathcal{R}=1)}{pr(\pi(\mathbf{X})=\pi, \mathcal{R}=1)} \\ &= \frac{pr(\mathcal{R}=1|\pi(\mathbf{X})=\pi, \mathbf{X}=\mathbf{x})\ pr(\mathbf{X}=\mathbf{x}|\pi(\mathbf{X})=\pi)\ pr(\pi(\mathbf{X})=\pi)}{pr(\mathcal{R}=1|\pi(\mathbf{X}=\pi))\ pr(\pi(\mathbf{X})=\pi)} \\ &= \frac{pr(\mathcal{R}=1|\pi(\mathbf{X})=\pi, \mathbf{X}=\mathbf{x}) pr(\mathbf{X}=\mathbf{x}|\pi(\mathbf{X})=\pi)}{pr(\mathcal{R}=1|\pi(\mathbf{X})=\pi)} \end{aligned}$$

Now

$$pr(\mathcal{R} = 1|\pi(\mathbf{X}) = \pi, \mathbf{X} = \mathbf{x}) = pr(\mathcal{R} = 1|\mathbf{X} = \mathbf{x}) = \pi(\mathbf{x}) = \pi$$

by definition of the propensity score and also the assumption that $\pi(\mathbf{x}) = \pi$ and by theorem 2 in (RR)

$$pr(\mathcal{R} = 1|\pi(\mathbf{X}) = \pi) = \pi$$

Thus

$$pr(\mathbf{X}=\mathbf{x}|\pi(\mathbf{X})=\pi,\mathcal{R}=1)=\frac{\pi pr(\mathbf{X}=\mathbf{x}|\pi(\mathbf{X})=\pi)}{\pi}$$
$$=pr(\mathbf{X}=\mathbf{x}|\pi(\mathbf{X})=\pi)$$

Similarly for $pr(\mathbf{X}=\mathbf{x}|\pi(\mathbf{x})=\pi,\mathcal{R}=0)$

$$pr(\mathbf{X}=\mathbf{x}|\pi(\mathbf{x})=\pi,\mathcal{R}=0)=\frac{pr(\mathcal{R}=0|\pi(\mathbf{X})=\pi,\mathbf{X}=\mathbf{x})pr(\mathbf{X}=\mathbf{x}|\pi(\mathbf{X})=\pi)}{pr(\mathcal{R}=0|\pi(\mathbf{X})=\pi)}$$

Now

$$pr(\mathcal{R}=0|\pi(\mathbf{X})=\pi,\mathbf{X}=\mathbf{x})=pr(\mathcal{R}=0|\mathbf{X}=\mathbf{x})=1-\pi(\mathbf{x})=1-\pi$$

by definition of the propensity score, the assumption that $\pi(\mathbf{x})=\pi$ and by theorem 2 in (RR)

$$pr(\mathcal{R}=0|\pi(\mathbf{X})=\pi)=1-\pi$$

$$\text{so } pr(\mathbf{X}=\mathbf{x}|\pi(\mathbf{x})=\pi,\mathcal{R}=0) = \frac{(1-\pi)pr(\mathbf{X}=\mathbf{x}|\pi(\mathbf{X})=\pi)}{(1-\pi)} \quad (4.2)$$
$$= pr(\mathbf{X}=\mathbf{x}|\pi(\mathbf{X})=\pi) \quad (4.3)$$

hence (4.1) □

From the above proposition the nearest neighbour obtained for the nonrespondent will have similar distribution of covariates of the respondent within the neighbourhood of $\hat{\pi}(\mathbf{X})$.

Use of propensity score for matching has two stages;

a Computation of propensity score

b Finding the nearest neighbour

4.6.1 Computation of Propensity Score

Unlike the sample selection probabilities, the probability to respond to an item in the questionnaire is not known a priori, hence these propensity scores are estimated from data by using either logistic regression, discriminant analysis or probit analysis. As most of the variables in our data are categorical we use logistic regression. As in RR, to estimate the propensity score we use the following form

$$\pi(\mathbf{X}) = \frac{e^{\mathbf{X}^*\beta}}{1 + e^{\mathbf{X}^*\beta}} \tag{4.4}$$

where $\mathbf{X}^* = (1, X_1, X_2, ..., X_v)$ and $\beta = (\beta_0, \beta_1, ...\beta_v)$. The *glm* function in 'R' software is used to estimate $\pi(\mathbf{X})$, by regressing the response indicator $\mathcal{R}$ against the covariates $\mathbf{X}$, with the family binomial and link function logit.

4.6.2 Nearest Neighbour by Propensity Scores (NNPS)

To determine the nearest neighbour, using propensity scores we have previously determined, we use the Euclidian distance measure. The distance measure is; $|\hat{\pi}(\mathbf{X}_i) - \hat{\pi}(\mathbf{X}_j)| \quad \forall i \in obs$ and $j \in mis$, where $\hat{\pi}(\mathbf{X}_i)$ is the estimate of $\pi(\mathbf{X}_i)$ obtained using equation (4.4). We use this distance measure because the covariate space is univariate and continuous. As $\hat{\pi}(\mathbf{X})$ is continuous we expect no ties among the donors obtained for imputing the missing values.

4.7 Conclusions

It is clear from the existing literature that use of all survey variables is not necessarily an ideal solution to the problem of donor imputation. A subset of survey variables can be selected by using professional judgement or by looking at the statistical association between the variables (Sixten and Sarndal, 2002; Westbrooke, and Jones,

2000). As a basis of comparison we use three different methods of data reduction, principal component analysis, graphical modelling and propensity matching for this thesis.

Principal component analysis is a very well known technique of data reduction in multivariate analysis. It has some advantages and limitations. The advantages are that instead of working on v variables one can use $w < v$ principal components. Using the techniques of Cadima and Jolliffe (2001), instead of choosing principal components which are linear combinations of variables we actually use a subset of the variables. This makes interpretations of the data easier for imputation. The limitations of principal component analysis are:

- The decision on the number of principal components or variables to be retained in a subset is not well defined.
- The subsets of variable may be a good representation of the covariates but their relation to the dependent variable may not be clear.
- Handling of data sets with mixed type of covariates is not known.

In view of the problems above of PCA and as another method we look at graphical modelling. Graphical modelling is a new approach which is now commonly being used for data reduction. This looks at the conditional independencies between the dependent variable and the covariates. We use this technique to take a subset of the variables. There are some advantages and problems in using this method. The advantages with this method are it handles both categorical and continuous variables easily and provides a subset which explains their relation with the dependent variable. It is a pictorial representation and hence can easily be interpreted. Unlike PCA, for graphical modelling the number of variables retained in the subset are obtained from the model once the conditional independence level is fixed (see 5.5.2).

With this method guideline ii) described in section 4.2.1 is satisfied. Even though this overcomes most of the problems of PCA, it is still not free from problems. For example

- For deleting the edge we use the edge deletion test multiple times, though each test is made at the nominal 5% significance level, the overall test has a much higher and unknown significance level (Whittaker, 1990).

- If a particular edge is deleted it is not known how to incorporate this information when deleting other edges.

For the subset obtained using these two methods we notice that only guideline ii) can be satisfied, but there is no guarantee that the guideline i) will be satisfied. With graphical modelling there may be a chance to capture the MAR property if the covariate that is responsible for nonresponse is correlated to the dependent variable, but with principal component analysis such relationship is not known. If guideline i) is not satisfied then the subsets will provide biased estimates. In real life situations there is no chance that one can know whether the missingness is MAR or not. Hence use of these methods for data reduction may not be very useful. Keeping this in mind we came up with a new method we call propensity matching.
Like graphical modelling, propensity matching can also handle both categorical as well as continuous variables. It is a sufficient summary of the covariates. As shown in (Prop 1), section 4.6, propensity score is a balanced score and by the results of Little and Rubin (2002, p-48) guideline i) holds for propensity matching. Even though this has advantage over the other two approaches there is still a limitation of this method in that, it is not clear whether guideline i) is satisfied or not. Suppose the covariate set has a covariate that is highly correlated with the dependent variable but not with the response indicator, then this covariate may or may

not be significant in the logistic regression used for finding the propensity scores. If this covariate is not significant in logistic regression then, according to Schafer [http://www.stat.psu.edu/ jls/mifaq.html], imputed values of the dependent variables may not have relationship with the covariates, thus distorting the joint distributions. However the marginal distributions are preserved.

From the above studies we conclude that only propensity matching can preserve the MAR condition when the multivariate data set of high dimension is reduced to a low dimension. Preserving this condition implies that nonresponse bias can be reduced. Hence data reduction by propensity matching may be useful for real life situations. In this thesis we tested all three methods and results are presented in chapter 5.

Chapter 5

Results and Discussion

5.1 Introduction

In chapter 3 we proposed a new method for imputing missing data when the data has both categorical as well as continuous covariates. This method of imputation has the advantage of being nonparametric and hence avoids the problems of model misspecification. In addition this method handles the categorical covariates without requiring transformations such as dummy variable creation to enable regression modelling. However while this method has advantages over other imputation methods it is computationally intensive. To overcome this complication we demonstrated in chapter 4 the use of propensity matching and two other methods for data reduction before imputation. Propensity score used in propensity matching is a balancing score and is the coarsest summary of the covariates (Rubin, 1985). The results described in this chapter illustrate the overall performance of our methods. Comparison is made in terms of its ability to predict the missing values close to the true values and preserving the marginal distribution. Various missing at random mechanisms for creating nonresponse was used in the simulations based on the NFHS-2 data.

Item nonresponse in NFHS-2 data was studied to develop response models that will be used for creating nonresponse in simulations. We describe the simulation process in section 5.3. We compare the performance of the new imputation method and a most commonly used method of imputation in section 5.4. Results from application of the data reduction methods are discussed in section 5.5. A comparison of the performance of MVNN method using the subsets obtained by covariance and correlation matrix in PCA is presented in section 5.6. Discussions on the results are presented in section 5.7.

5.2 Item Nonresponse in NFHS-2

In this thesis we look at item nonresponse in the hemoglobin level (HL) variable. The main reason for looking at this particular item is that the NFHS-2 survey, in the state of Uttar Pradesh (U.P) has an overall nonresponse of 38% for HL. In addition, in NFHS-2 survey the testing of hemoglobin was conducted for the special purpose of studying maternal mortality due to anaemia, anaemia being a low level of haemoglobin in blood (Shilpa Sapre, 2001).
The details of nonresponse for hemoglobin in NFHS-2 data are as follows:

- Item nonresponse by region show that the nonresponse was concentrated in region 1 and region 5 (Appendix C).

- The rural, urban breakdown of nonresponse shows that the rural respondents did slightly better than their urban counterparts. The percentage of nonresponse in the rural area was approximately 38%, whereas for urban areas it was approximately 40%.

- For religion covariate, it was observed that Muslim women had higher nonre-

sponse compared to other religions.

- Similar observations are made for nonresponse distribution in other covariates such as age, standard of living etc.

For adjusting the nonresponse in HL we propose stochastic single imputation. To recommend any particular method for imputing the nonresponse we need to have knowledge on the performance of the imputation. To do this on real data would be difficult as the values for nonrespondents by definition are not known. Some researchers have endeavored to collect their values via follow-up surveys but there is always some nonresponse remaining. For much imputation research simulations studies are done. In simulations one has the advantage of knowing the true values of missing data, and so can assess the performance of the imputation method.

5.3 Simulations

Rather than using fictitious data for our simulation study, the complete set of respondents in NFHS-2 were mainly used as the fully observed (i.e. complete case) population.

5.3.1 Details of Simulation Population

The original data of NFHS-2 had more than 2000 variables (dependent and covariates) spread over several sections. From these set of sections we selected a set of covariates. Selection of relevant sections is described in tables 4.1-4.2, p.68-69. The data set thus selected for simulations has 29 variables and the list of covariates is presented in Appendix B. In simulations where we compare MVNN and RBNN imputation methods under simple MAR, we selected sample sizes of 300, 500, and

1000, from the fully observed population. It is only in simple MAR we use sample sizes of 300, 500 and 1000. Since MVNN was computationally intensive, for other comparisons of MAR models we used 300 and 500 sample sizes. For this study we ignored the multistage sample design of NFHS-2 and selected the samples using simple random sampling without replacement design (SRSWOR). We used SRSWOR because these computations can often be used as a convenient base for comparing the results obtained with more complex designs (Kish, 1965). Complex design needs to be allowed when estimating the extra variance due to imputation. This we have investigated in chapter 6. Moreover the primary motive is to compare the two imputation methods.

These samples which are drawn from the population are initially fully observed. Before creating nonresponse in the data, the samples thus selected were duplicated so that the true values for the nonrespondents in the samples can always be known. Nonresponse was created for the experiments using each one of the MAR models outlined in chapter 3. All these response models are defined in sec 3.8. Figure-5.1 shows the simulation process. In the simulations we used various levels of nonresponse. Even though the NFHS-2, nonresponse is 38%, in our simulations we use a range of nonresponse levels $(5\%, 10\%, 15\%, 25\%)$. These values allows us to investigate and compare how our methodology works for various nonresponse levels and with various sample sizes. Imputation was carried out on data sets thus created and the imputed data are compared to the fully observed sampled data.

5.4 Comparison of Imputation Procedures

In this section we present the results obtained for the data imputed using both regression based nearest neighbour (RBNN) and multivariate nearest neighbour (MVNN)

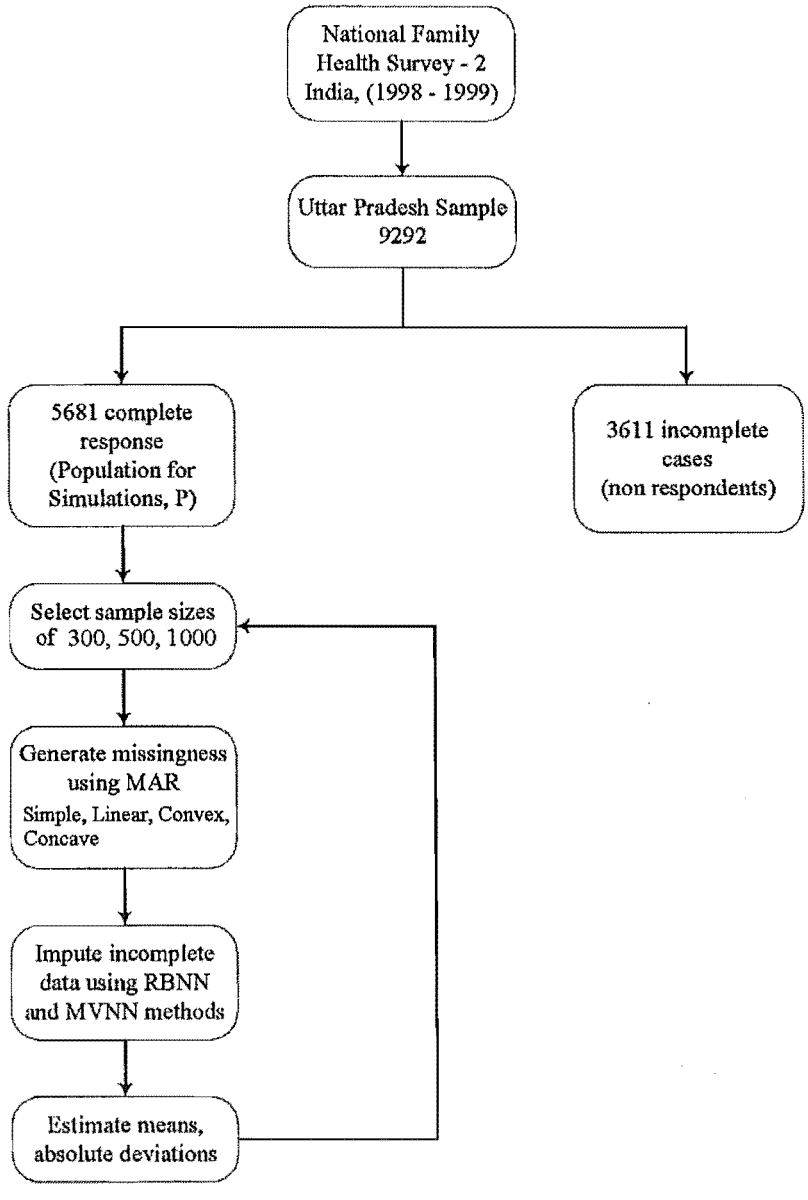

Figure 5.1: Schematic representation of the simulation process

imputation methods. For applying the RBNN and MVNN methods on the samples described in section 5.3.1, an R program was used for writing the code for both the methods. This code can be used for any real data with nonresponse. Both RBNN and MVNN methods are studied under different nonresponse rates, different sample sizes and response models. Their performance is compared using four measures, mean square error of the parameter (MSEM), mean square error of the imputed values (MSEI), mean absolute deviations (MAD) all these measures describe imputing the values close to the true values. Leti's index is to study the homogeneity of distributions.

5.4.1 RBNN Simulations

For the regression part of RBNN method we used linear model package in R. In this method the dependent variable Y was HL and the covariate $\mathbf{X}$ was a matrix of dimension $n \times (v = 28)$ (see Appendix B). As in earlier chapters of this thesis we assume nonresponse in one variable (HL), hence $HL = (HL_{obs}, HL_{mis})$. The corresponding covariates $(\mathbf{X})$ which are completely observed are $\mathbf{X}_{obs}, \mathbf{X}_{mis}$. Even though some of the covariates were not significant in the regression we still used them as they were expected to be among the possible factors that explain the variation in HL. Rubin (1976) has recommended the use of all covariates to the fullest extent because they help reduce the bias, the covariates may be of importance to the subject related experts. For performing the regression-based nearest neighbour imputation, the following model was used:

$$E(HL_{obs}) = \mathbf{X}^*_{obs}\beta \tag{5.1}$$

where $\mathbf{X}^*_{obs} = (1, \mathbf{X}_{obs})$ and $\beta = (\beta_0, \beta_1, ...\beta_v)$ with β_0 being the intercept. The categorical covariates were appropriately transformed into dummy variables before

using them in the regression. As explained in section 3.3 the estimated coefficients of β, $\hat{\beta}$, were used in predicting entire HL $(\widehat{HL})$=$(\widehat{HL}_{obs}, \widehat{HL}_{mis})$. That is

$$\widehat{HL} = \mathbf{X}^*\hat{\beta} + \epsilon.$$

where $\mathbf{X}^* = (1, \mathbf{X})$ and ϵ is the error term. This error term $\epsilon \sim N(0, \sigma^2)$ and σ^2 is the residual mean square error of the (5.1). The $\widehat{HL}_{obs}$ and $\widehat{HL}_{mis}$ are used for finding the nearest neighbour (NN). As $\widehat{HL}$ is a continuous variable we use Euclidean distance measure in finding the nearest neighbour, defined as

$$d_{ij} = |\widehat{HL}_i - \widehat{HL}_j| \ \forall i \in obs, j \in mis.$$

To obtain the nearest neighbour to provide a donor for the missing case j, that is the case k where k : d_{kj} =$\min_{1 \leq i < r}(d_{ij})$, we use the HL_{obs} corresponding to the nearest neighbour obtained as the imputed value for HL_{mis}. For the RBNN model the R^2 ranged between 0.25 and 0.72 over the 1000 simulations for each sample size of 300, 500 and 1000.

5.4.2 MVNN Simulations

To enable us to compare the performance the simulation data used for MVNN is same as that used in RBNN imputation. To find the nearest neighbour using MVNN we compute the dissimilarities between the responding case c and the nonresponding case m. These dissimilarities were computed using the equation 3.3. From this equation it is seen that the dissimilarity is the weighted sum of the distances computed between the cases using all covariates. In the computer program we specify the type of variable. The function type of variable then helps in recognizing the

form of distance measure d^j_{cm} to be used in the dissimilarity computation. Here

$$d^j_{cm} = \begin{cases} \left.\begin{array}{ll} 1 & \text{if } x_{cj} \neq x_{mj} \\ 0 & \text{otherwise} \end{array}\right\} & \text{for binary and nominal} \\ \frac{|x_{cj} - x_{mj}|}{r_j} & \text{interval and ordinal variables} \end{cases}$$

When the covariate is asymmetric binary as defined in section 3.4 and $x_{cj} = x_{mj} = 0$ then δ^j_{cm} in eq.(3.3)=0. This computation of dissimilarity is repeated for each case in the data, this leads to a $n \times n$ matrix of dissimilarities. This dissimilarity matrix is a symmetric matrix. The diagonal elements of the dissimilarity matrix are 0 (since the dissimilarity for self is 0). There is a chance that some of the off diagonal elements will be zero. This indicates that there are certain cases which are identical to the missing case which, of course, is a situation preferred in matching, as we take the case with the minimum dissimilar value as the donor. The HL_{obs} value of the donor so obtained is used for imputing the missing value. Sometimes there may be more than one donor, that is, there are several candidate cases with same dissimilar value. In such situations, we randomly select a donor from among these candidates.

5.4.3 Results

The performance of RBNN and MVNN is compared using the four measures defined in section 3.9. For the mean square error of the parameter measure, we use the mean $f(Y) = \bar{Y}$ (totals can be obtained by multiplying the means by the sample size). Both the methods are compared using the four response models described in section 5.3.1.

5.4.3.1 Simple MAR Mechanism

This is the most commonly assumed response model in the literature on nonresponse. Here the nonresponse in HL can be explained by one or more of the covariates. In

order to generate a simple MAR mechanism we use two covariates.

We compare the performance of MVNN and RBNN using three different sample sizes because the increase in sample size increases the probability of finding a donor very similar to the nonrespondent as we have a larger pool of possible donors. Three different nonresponse rates were used. Each of these combinations were simulated 400 times. Initial results were encouraging and are reported in Murthy *et al* (2003) and are given in Appendix -E. These results motivated us to further investigate the performance of MVNN method. For these further studies, we used religion and current pregnancy status as the covariates which model nonresponse, that is, we set nonresponse in hemoglobin (HL) if

$$[((\text{religion} = \text{Muslim}) \cap (\text{pregnancy status} = \text{currently pregnant})) \cup (U < \vartheta)]$$

where U is an independent random variable uniformly distributed over [0,1] and ϑ is a constant that is used to modify the probability that a case has missing value. It is from our experience in NFHS-2 on what causes nonresponse (section 2.1.2) that led us to use religion as a nonresponse covariate. Using these two situations, the performance of RBNN and MVNN are compared for different nonresponse rates and sample sizes. The performance of MVNN and RBNN were assessed using four different measures MSEM, MSEI, MAD and Leti's index as explained in section 5.4. The results obtained under this simple response model, given in Table-5.1 and Figures 5.2 - 5.7 are explained below.

Mean square error of means (MSEM): As described in previous section, the parameter of interest in this study is mean level of hemoglobin ($\bar{HL}$). Hence for the first comparisons we looked at the difference in the means obtained for HL from imputed data and the true sample. In order to keep the units of comparison (grams/deciliter) same across the comparisons we take the square

Table 5.1: Comparisons of the performance of the data imputed by MVNN and RBNN methods under simple MAR nonresponse, range of HL [70-170(g/dl)], population mean=118, sd=19.07

Sample Size 300									
Method	Nonresponse								
	5%			10%			15%		
	MSEM	MSEI(se[a])	MAD	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD
MVNN	0.62	26.1(0.4)	20.94	0.78	25.2(0.2)	19.91	0.65	25.3(0.2)	18.69
RBNN	0.32	27.6(0.6)	21.12	0.50	26.7(0.4)	20.38	0.68	26.6(0.2)	19.89
Sample Size 500									
MVNN	0.17	25.7(0.3)	20.08	0.39	24.4(0.1)	19.05	0.37	24.1(0.2)	19.59
RBNN	0.26	28.1(0.4)	21.77	0.44	26.6(0.3)	20.91	0.56	26.6(0.2)	20.96
Sample Size 1000									
MVNN	0.17	24.8(0.2)	19.44	0.36	22.3(0.1)	18.67	0.26	22.5(0.0)	18.70
RBNN	0.17	26.6(0.2)	20.63	0.24	25.3(0.3)	19.83	0.35	25.1(0.1)	20.14

[a]standard error

root of the mean square error obtained using (3.9). The final result presented in table-5.1 is the root mean square error obtained over all simulations. We notice that as the nonresponse rates increase the MSEM usually increases, reflecting the natural loss of information that occurs with higher rates of missing data in HL. When the sample size is increased, as expected, the MSEM decreases. With the increase in sample size the chances of finding a donor close to the characteristics of the nonrespondents increases. For both methods the bias is fairly large when the sample size is 300 but as the sample size increase the bias is reduced

From the box plots presented in Figure(5.2-5.4) it is observed that across the nonresponse rates the distribution of bias for MVNN is consistently small compared to RBNN. Also the inter quartile range and the spread of the distribution of bias for MVNN is generally smaller compared to RBNN.

Mean square error of the imputed values: The second measure that we use in this study measures the closeness of the imputed values to the real values. This is the MSEI along with their standard errors, presented in Table-5.1 indicate that, it is consistently smaller for MVNN over all the nonresponse levels and sample sizes compared to that for RBNN. These differences if taken in conjunction with the standard errors indicate that the differences between both MVNN and RBNN are statistically significant. On two occasions the MSEI for MVNN (RBNN does the same) increased slightly for 15% missingness compared to 10%. This increase of MSEI for both the methods is not statistically significant. From the results of MSEI we can infer that MVNN consistently has smaller MSEI than RBNN.

Mean Absolute deviations: As outlined in chapter 3, the mean absolute deviations

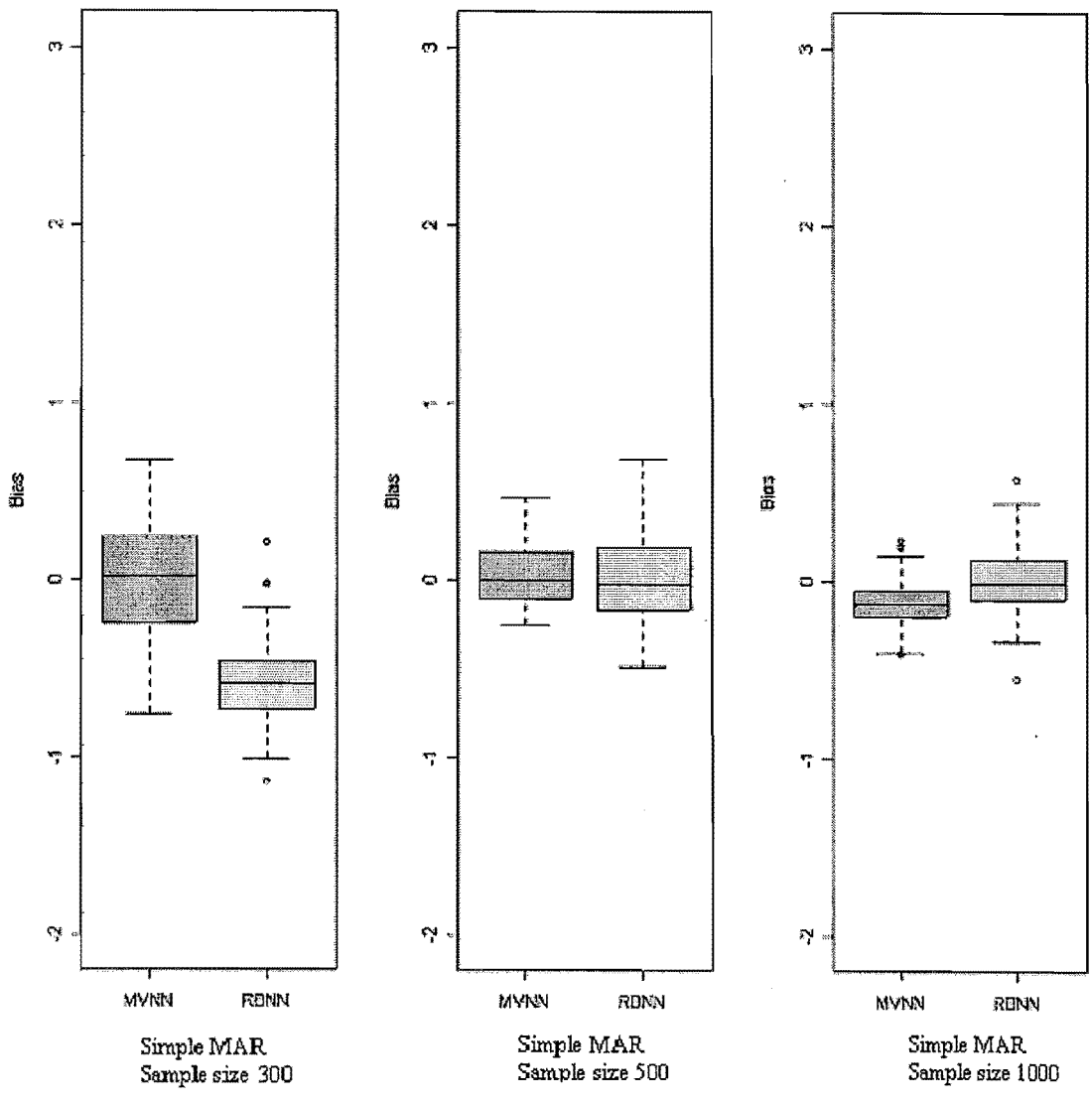

Figure 5.2: Distribution of bias for 5% nonresponse

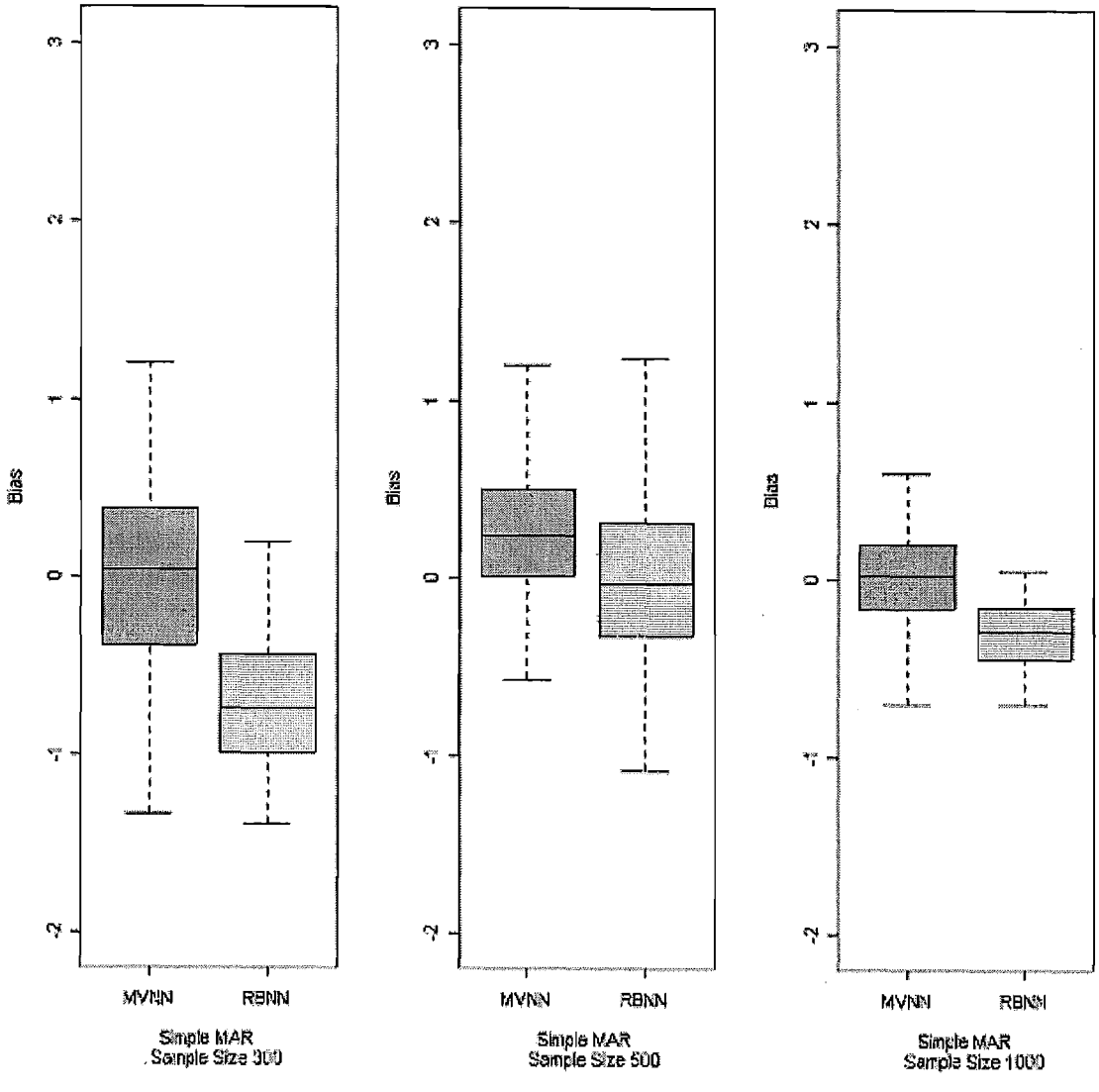

Figure 5.3: Distribution of bias for 10% nonresponse

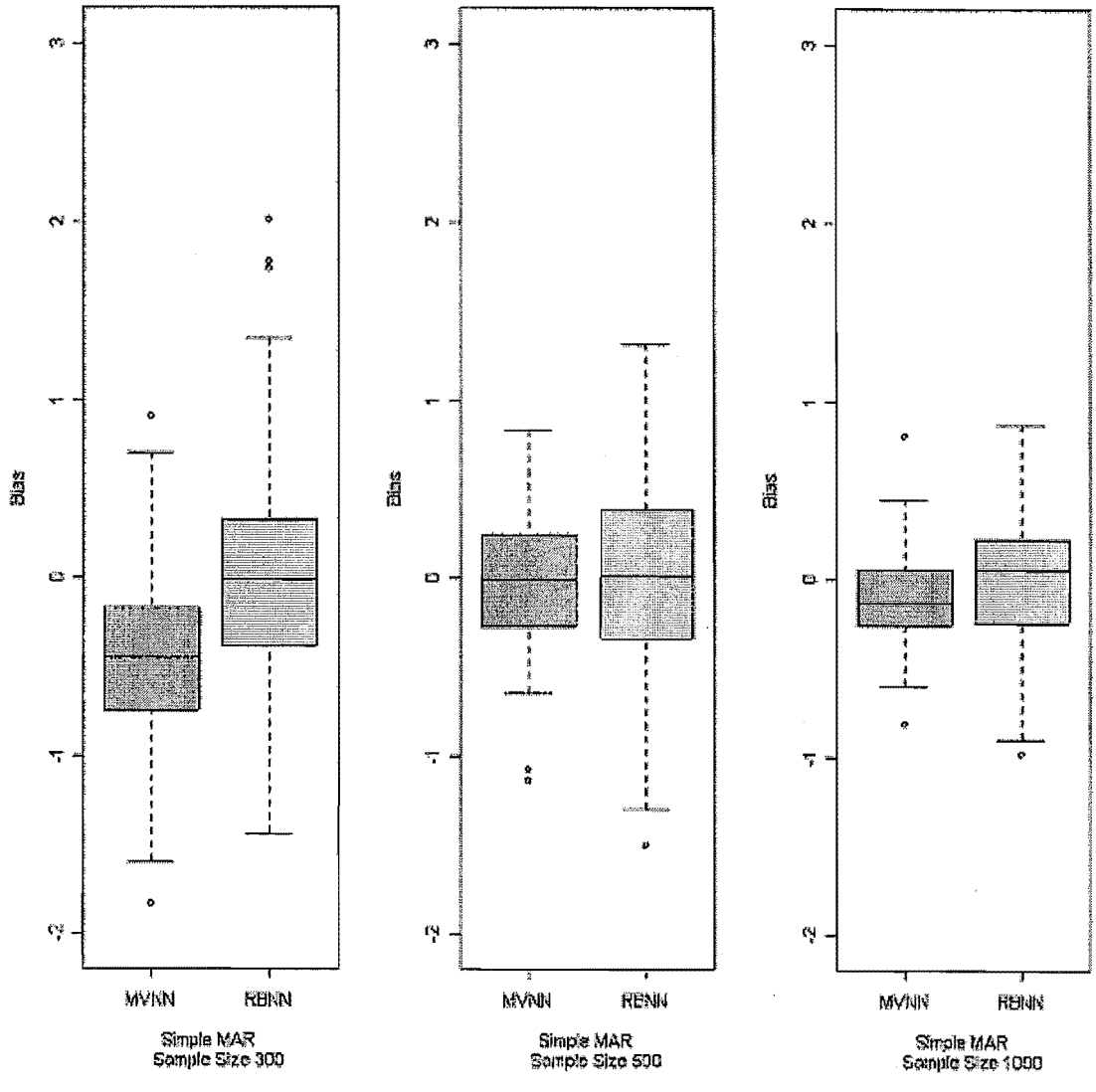

Figure 5.4: Distribution of bias for 15% nonresponse

are computed as the average of the absolute difference between the imputed value and the true value. The mean absolute deviations are also smaller for MVNN compared to RBNN, indicating that MVNN consistently imputes non-response closer to the true values.

Imputed marginal distributions To compare the relative distributions of the data imputed using both MVNN and RBNN we used Leti's index as defined in section 3.9.3. As described in section 3.9.3 the index can have values between 0 to 100. A low value shows that the distributions of each category of HL in the imputed data and the true data are similar. As the dependent variable (HL) was continuous it was categorized for the purpose of calculating the Leti's index. This categorization was done on the basis of the severity of anemia. The variable HL thus categorized has four categories, namely "Severe" (less than 70 g/dl), "Moderate" (70-99 g/dl), "Mild" (100-109 g/dl), and "Normal" (110+ g/dl), a classification used in NFHS-2 data collection (IIPS, 2000). As there are only 4 categories it is not a strong distributional test but still should give an indication as to the preservation of the marginal distributions. Box plots of Leti's index for MVNN and RBNN methods are presented in Figure (5.5-5.7). From the box plots it is observed that distribution of HL is better preserved when the sample sizes are large, which could be expected given the increased possibility of finding more matches close to the characteristics of the nonrespondents. As the sample sizes are increasing the spread in the distribution of index is decreasing for MVNN. This again confirms the findings from other measures that MVNN imputes values close to the true values.

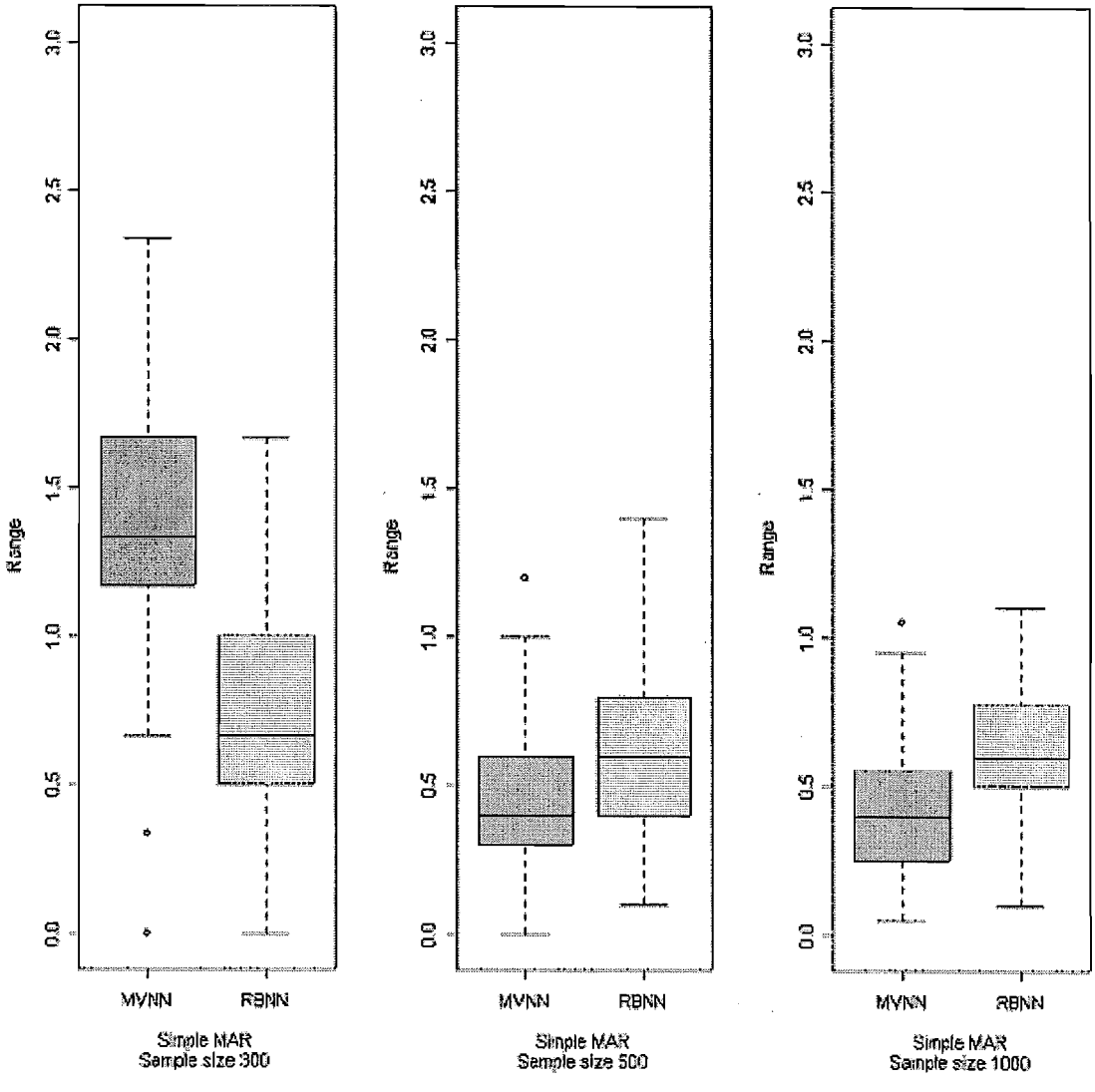

Figure 5.5: Distribution of Leti's index for 5% nonresponse

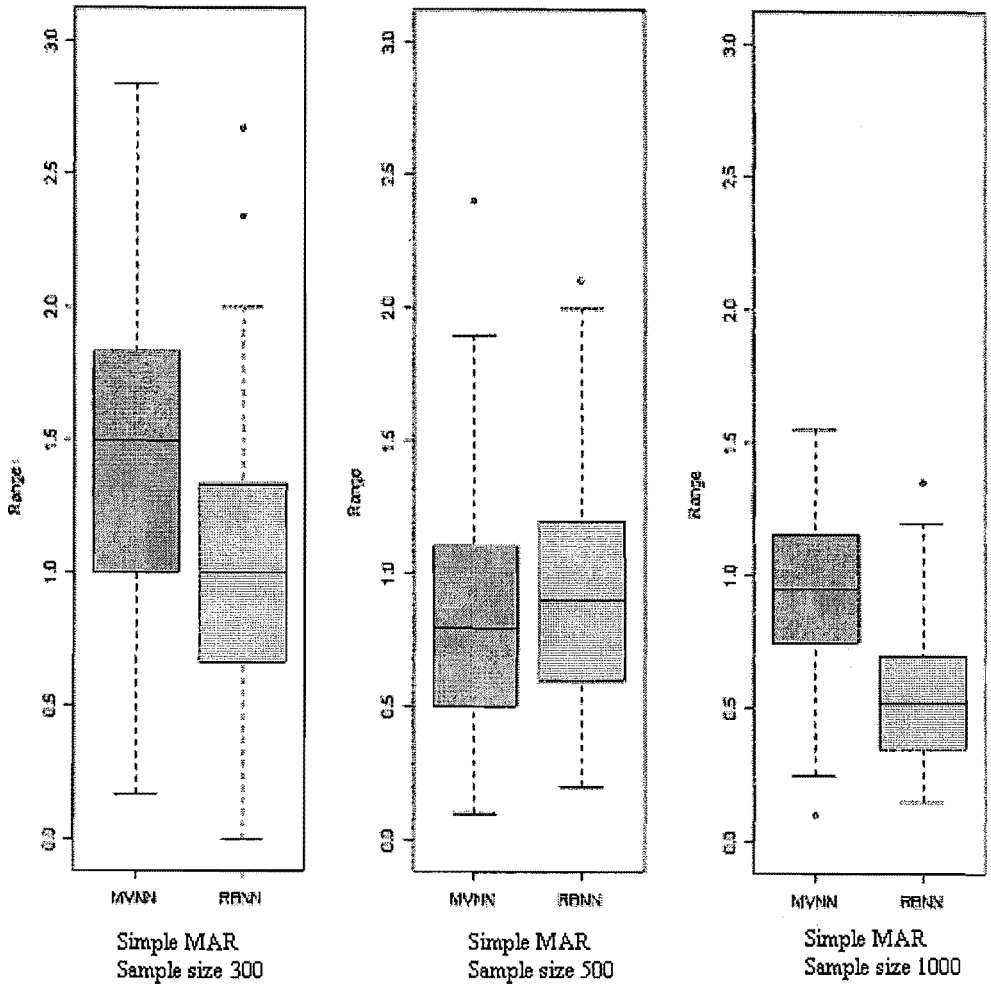

Figure 5.6: Distribution of Leti's index for 10% nonresponse

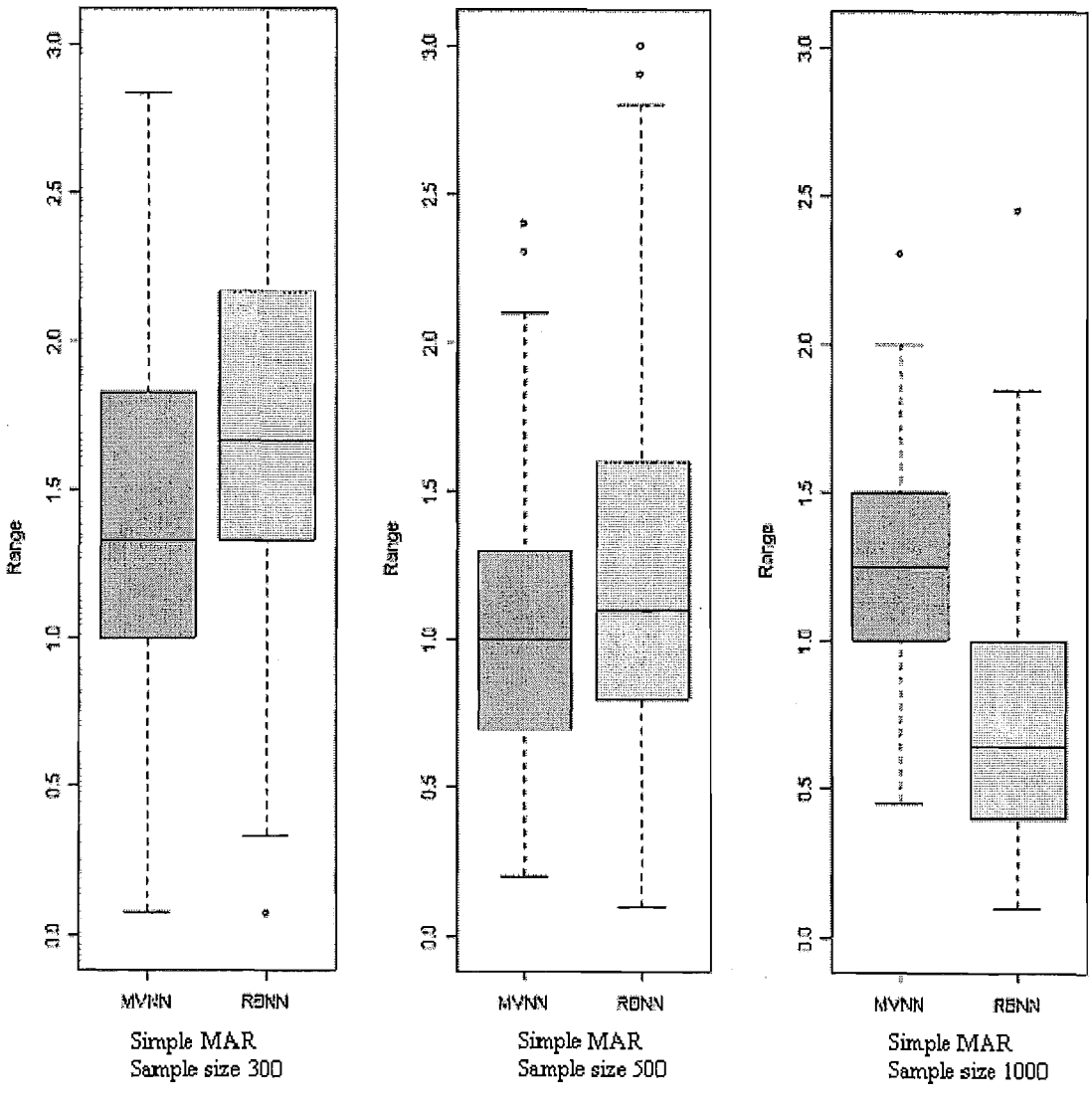

Figure 5.7: Distribution of Leti's index for 15% nonresponse

5.4.3.2 Overall Summary of MVNN and RBNN for Simple MAR

From the above results it can be seen that as expected the bias increased as the sample size decrease and or the nonresponse rate increase. The distribution of bias presented in box plots indicate that MVNN has a smaller spread in its distribution, indicating the consistency of imputations and a higher probability of lower bias for any single imputation. However in this case all the biases are small relative to the population mean of 118 (g/dl) of HL.

The MSEI values for MVNN do not lie within the 2 standard errors of the MSEI of RBNN indicating that, the differences are significant. This indicates that the MVNN imputes values closer to the true values. Our studies motivated us to compare the performance of MVNN and RBNN with various other MAR response models.

5.4.4 Additional Response Models Tested

Since it is likely that there are varying patterns of MAR nonresponse in surveys, we investigate variants of MAR, specifically MAR linear, MAR Convex and MAR Concave as defined in section 3.8. For these we used the covariate children ever born (CEB) to create nonresponse in HL. In all examples we divide CEB into four classes (< 2, 2-4, 4-6, 6+) and assigned nonresponse probabilities to each class.

- For MAR linear we used the probabilities (0.1, 0.2, 0.3, 0.4) for 15 percent nonresponse, and (0.2, 0.4, 0.6, 0.8) for 25 percent nonresponse.
- For MAR Convex we used (0.1, 0.2, 0.2, 0.1) for 15 percent nonresponse and (0.2, 0.4, 0.4, 0.2) for 25 percent nonresponse.
- For MAR concave we used (0.2, 0.1, 0.1, 0.2) for 15 percent nonresponse and (0.4, 0.1, 0.2, 0.4) for 25 percent nonresponse.

Nonresponse rates of 15% is often accepted as a reasonable nonresponse and any nonresponse above 25% is considered as high nonresponse rate.

5.4.4.1 Imputation for 15% Nonresponse

Table-5.2 presents the mean square errors of the mean, mean square error of the imputed values, and the mean absolute deviations, for MAR linear, MAR convex and MAR concave, obtained from the data imputed using MVNN and RBNN methods. Once again for this comparisons we used the same data sets with sample sizes 300 and 500 for both MVNN and RBNN. Since for a sample size of 1000 the MVNN method was computationally intensive hence from hereafter in our comparisons we use sample sizes of 300 and 500. As in simple MAR the differences in MVNN and RBNN methods for MAR linear, convex and concave are not distinguishable. One possible reason may be that as MAR is linear, convex or concave the nonresponse in HL is higher for some subpopulation this may restrict the set of possible number of donors for MVNN method thus increasing the MSEI and MAD. The general observation from the results is:

MSEM: Once again the MSEM presented in the table is the root mean square error (RMSEM). Except for MAR linear as the sample sizes increases the biases decrease, indicating an increased probability of finding a donor close to the characteristics of the nonrespondent. To show the distribution of bias over the simulations we present the box plots for the bias obtained (Fig.5.8). From the box plots we note that there are many more outliers for MAR linear compared to other mechanisms. This shows a possibility that the average presented in the table may be effected by these outlier leading to the high values of MSEM.

Table 5.2: Comparison of the performance of MVNN and RBNN method under MAR linear, MAR convex, MAR concave, and data with 15% nonresponse

Sample Size 300									
	Type of missing								
Method	Linear			Convex			Concave		
	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD
MVNN	0.46	26.6(0.1)	20.77	0.52	25.8(0.1)	20.54	0.58	26.7(0.1)	20.72
RBNN	0.46	26.8(0.1)	20.93	0.58	26.5(0.1)	20.84	0.65	26.9(0.1)	22.20
Sample Size 500									
MVNN	0.51	26.5(0.1)	20.65	0.48	26.7(0.1)	20.60	0.50	27.0(0.1)	20.85
RBNN	0.52	26.8(0.1)	20.91	0.50	26.6(0.1)	20.62	0.42	27.3(0.1)	21.20

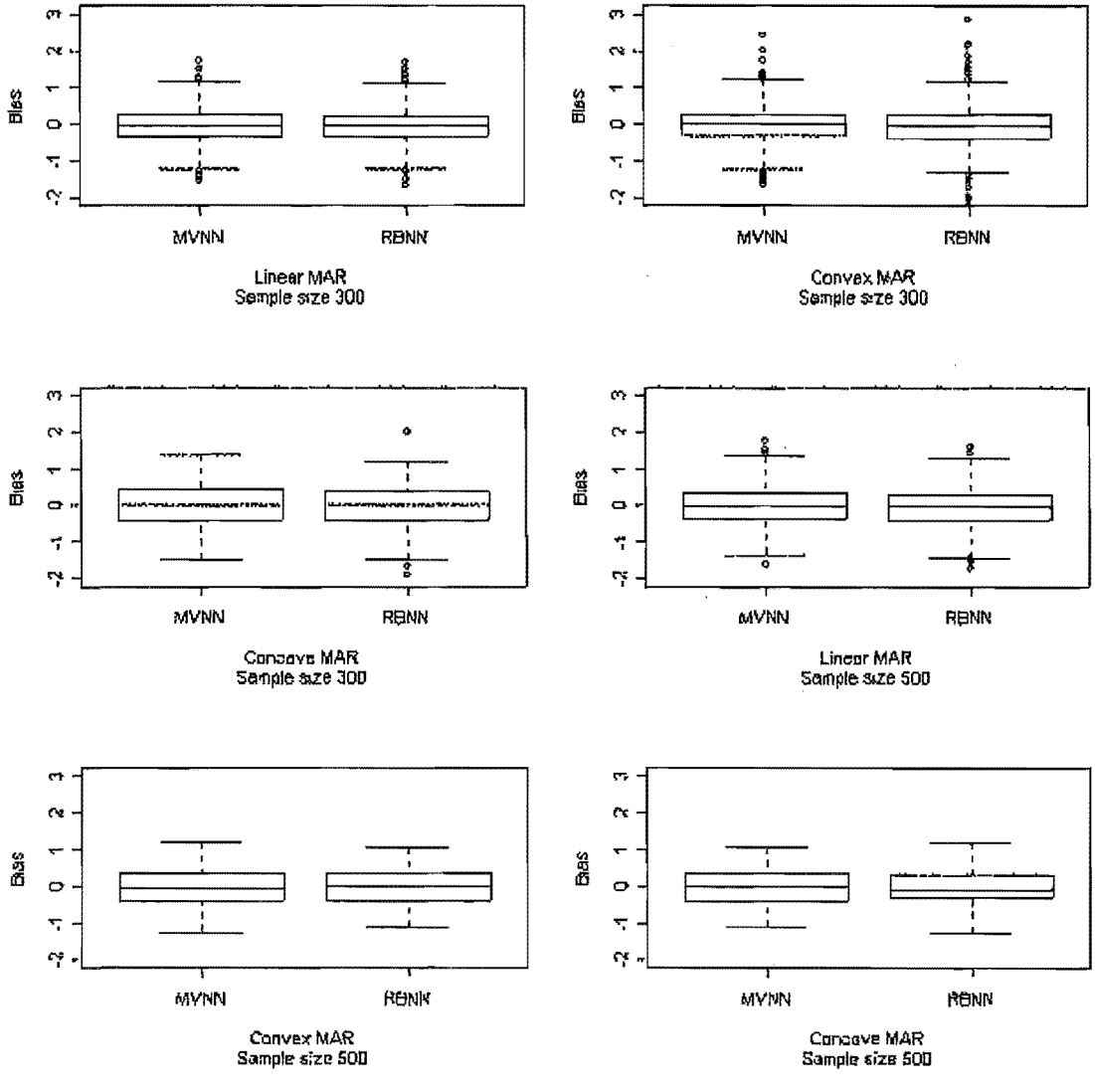

Figure 5.8: Distribution of bias for 15% Nonresponse

MSEI: To study the differences in the individual values of the imputed data we use MSEI. The values presented in the table are the root mean squared error of the imputed value. The MSEI of MVNN lie within the 2 standard deviations of MSEI of RBNN, hence the differences between the methods are not statistically significant. It is only for MAR convex we found a slight statistical significance for sample sizes 300, but we would not say that this is general as the differences disappear when sample size is 500.

MAD: The differences are not significantly different, but overall MVNN has smaller values than RBNN.

Leti's index From the box plots presented in Figure-5.9 we see that with small nonresponse rates and sample sizes MVNN consistently preserves the relative distributions of HL in the classes (severe, mild, moderate, normal) better than RBNN.

5.4.4.2 Imputation for 25% nonresponse

Table-5.3 shows MSEM, MSEI and MAD for the two imputation mechanisms. For these simulations we notice that as the nonresponse increase, the bias also increases. This is observed for both the imputation methods and for both sample sizes. With an increase in the sample size there is a reduction in the bias but this reduction is not very large. From the box plots presented in figure-5.10 we see that both MVNN and RBNN have similar distribution in bias, but the interquartile range for RBNN is slightly more than that of MVNN, indicating that over the repeated simulations, MVNN is more likely to impute the missing values more correctly than RBNN method. As sample size increases we notice the spread of bias in MVNN is much less compared to RBNN. As in previous sections the differences in MSEI

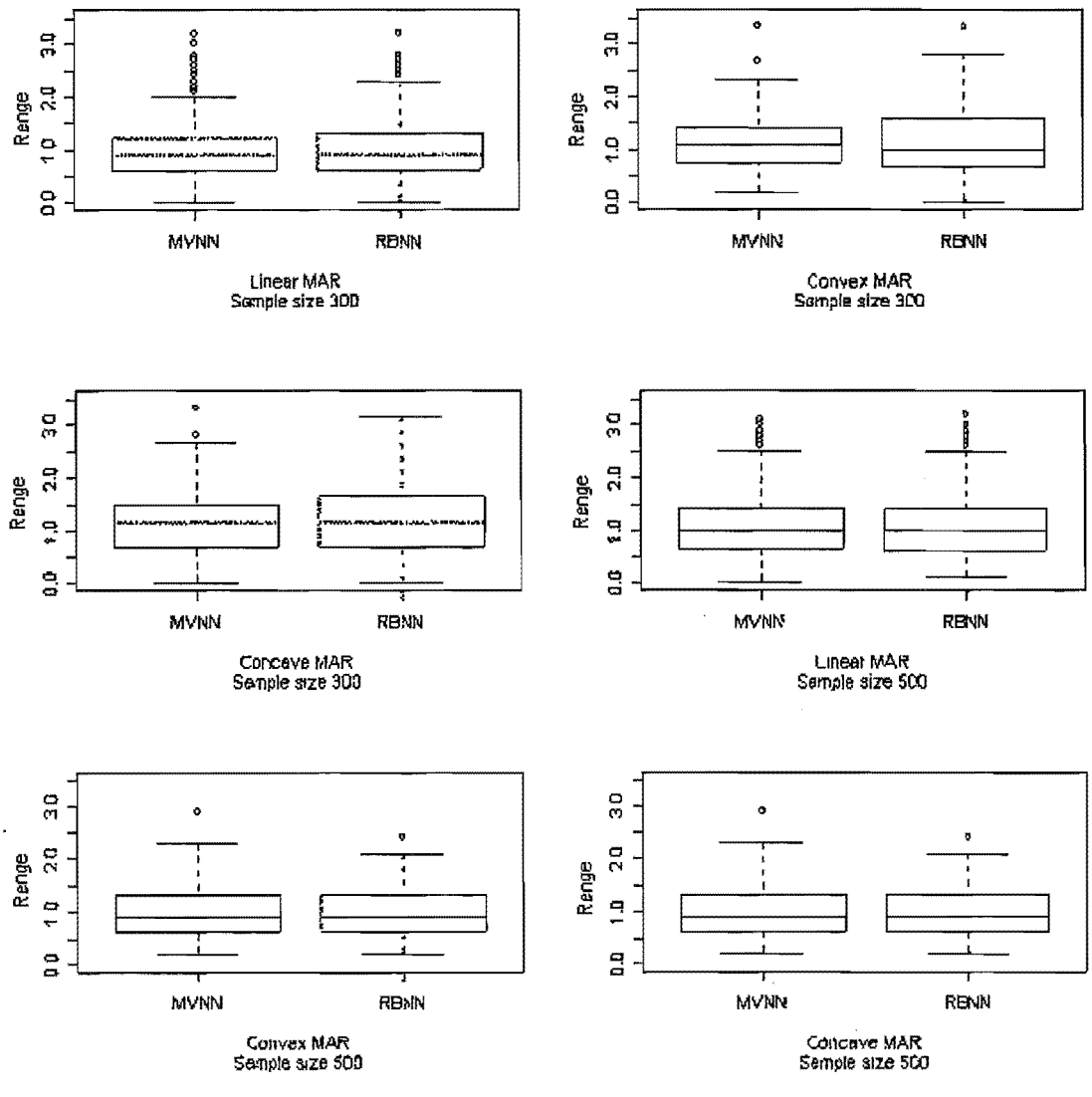

Figure 5.9: Distribution of Leti's index for 15% nonresponse

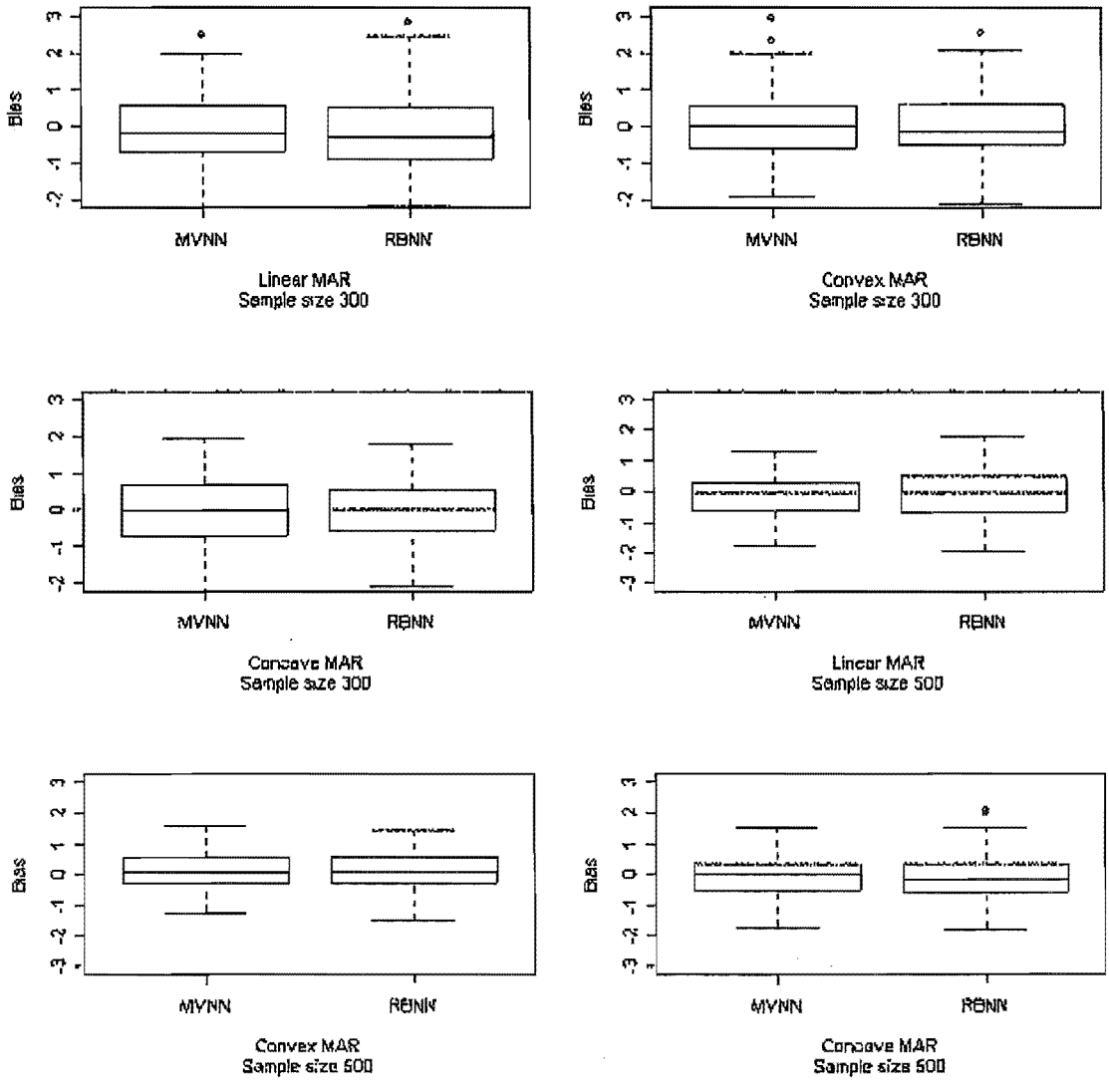

Figure 5.10: Distribution of bias for 25% nonresponse

Table 5.3: Comparison of the performance of MVNN and RBNN method under MAR linear, MAR convex, MAR concave, and data with 25% nonresponse

Sample Size 300									
	Type of missing								
Method	Linear			Convex			Concave		
	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD
MVNN	0.91	26.7(0.1)	20.70	0.95	26.3(0.1)	20.64	0.92	26.4(0.1)	20.65
RBNN	1.05	26.9(0.1)	20.81	0.98	26.6(0.1)	20.89	0.90	26.2(0.1)	20.66
Sample size 500									
MVNN	0.67	26.8(0.1)	20.87	0.65	26.6(0.1)	20.75	0.68	26.6(0.1)	20.55
RBNN	0.72	26.8(0.1)	20.91	0.65	26.8(0.1)	21.01	0.68	26.9(0.1)	20.89

are not statistically significant. The MAD is consistently smaller for MVNN as the sample size increases, indicating that the imputed values obtained using MVNN are generally closer to the true values than that for the RBNN method. From the figure-5.11, we observe that for sample sizes 300 and 500 and MAR convex and concave models, the distributions of Leti's index is left skewed for both MVNN and RBNN. This indicates that both MVNN and RBNN have many low index values, indicating that both generally preserve the distribution of HL.

5.4.5 Summary of MVNN and RBNN Comparisons

Even though the differences between MVNN and RBNN are not statistically significantly different in many cases, it should be noted that MVNN usually performs better than RBNN and is not likely to be worse than RBNN. This coupled with

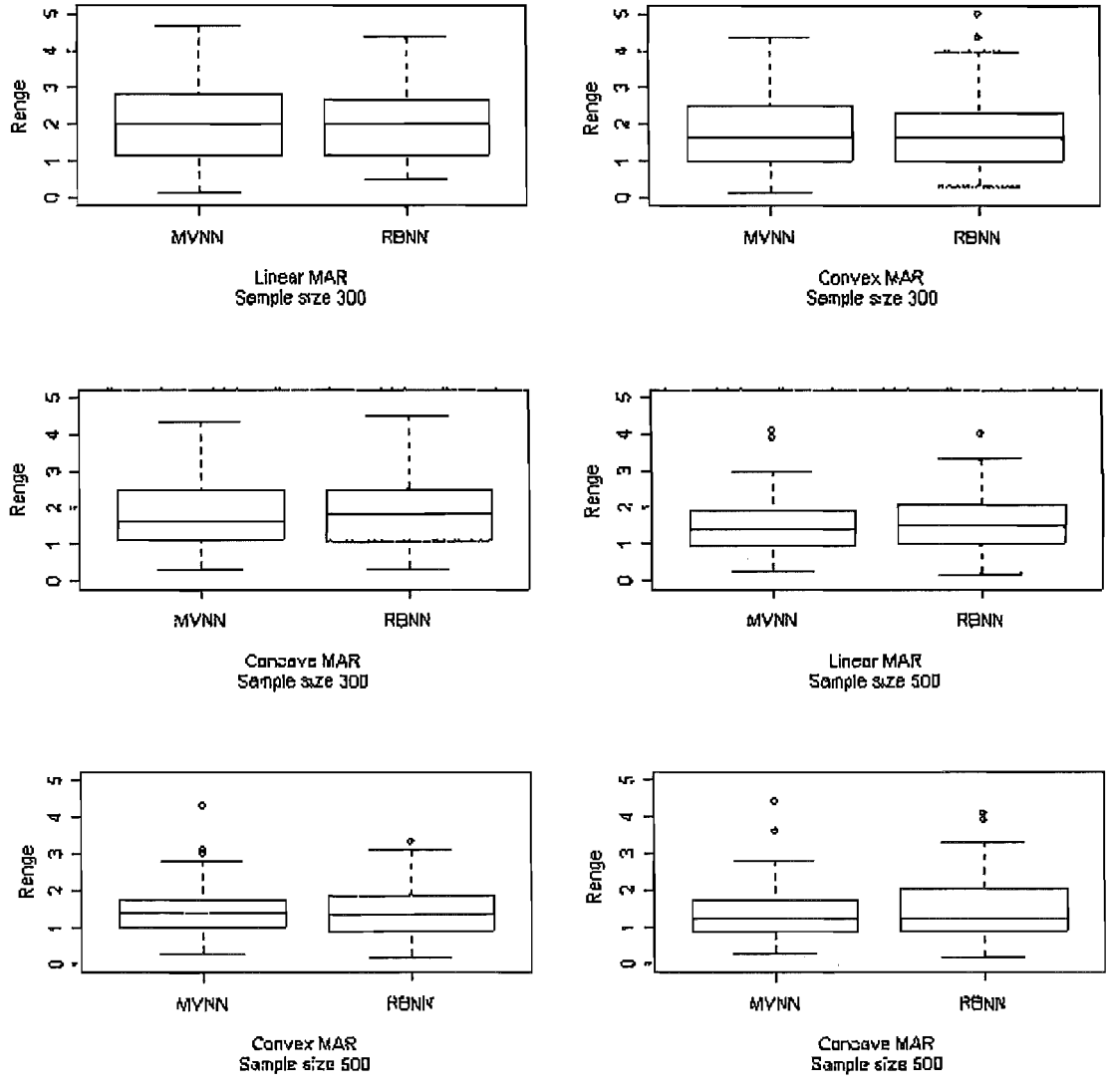

Figure 5.11: Distribution of Leti's index for 25% nonresponse

the other advantages of MVNN make it good method for imputation. However as we have noted MVNN is computational intensive. Thus we need to look at data reduction methods which create a subset of covariates, and then we can use MVNN method on this subset of covariates to impute the missing values.

5.5 Comparisons of Data Reduction Methods

As again we assume that the covariates are completely observed, nonresponse is in HL and that nonresponse is created using one of the response models, MAR simple, MAR linear, MAR convex and MAR concave. "Children ever born" (CEB) covariate was used to create nonresponse. Similar to the earlier comparisons presented in above sections we evaluated the performance of MVNN using the full set of covariates compared to using subset of covariates using sample of sizes 300 and 500 and for 5%, 10% and 15% percent nonresponse. From Table-5.6 we see that, there are no significant differences between the data imputed using subset of covariates and the full set of covariates. Hence in other MAR linear mechanisms we used sample sizes of 300 and 500 with 15% and 25% nonresponse rates. All these combinations were tested using 1000 simulations. For our simulation work data reduction was performed using propensity matching, graphical modelling and principal component analysis. As expected MVNN is computationally much quicker imputing using the subsets obtained by these methods.

5.5.1 Propensity Matching

As described in section 4.6, the covariate set can be reduced to a single propensity score, which is a complete summary of the covariates. As outlined in section 4.6, these propensities can be estimated from the sample data using logistic regression.

For computing the propensity score, the "glm" function in R was used. In this glm function, the family was set to binomial and the link function used was logit. The dependent variable in the "glm" function is the response indicator ($\mathcal{R}$) and the independent variables are the set of covariates ($\mathbf{X}$). Hence the model for finding the propensity score

$$\pi(\mathbf{X}) = pr(\mathcal{R} = 1|\mathbf{X}) = \frac{e^{\mathbf{X}^*\beta}}{1 + e^{\mathbf{X}^*\beta}} \tag{5.2}$$

where $\mathbf{X}^* = (1, \mathbf{X})$ (covariate list in Appendix B) and $\beta = (\beta_0, \beta_1, ...\beta_v)$. In logistic regression the categorical covariates were appropriately transformed to dummy variables. Let $\hat{\pi}(\mathbf{X}_i)$ be the estimated propensity score for i^{th} observed cases in the sample and $\hat{\pi}(\mathbf{X}_j)$ be the estimated propensity score for the j^{th} nonrespondent in the sample. To find a nearest neighbour for the j^{th} nonrespondent from the pool of possible donors, because the propensity scores are continuous we use the Euclidian distance measure.

$$d_{ij} = |\hat{\pi}(\mathbf{X}_i) - \hat{\pi}(\mathbf{X}_j)| \quad \forall i \in obs \text{ and } j \in mis$$

where $\hat{\pi}(\mathbf{X}_i)$ is the estimate of $\pi(\mathbf{X}_i)$ obtained using equation (5.2). To obtain the nearest neighbour for the missing case j, we choose the case k where k : $d_{kj} = \min_{1 \leq i < r}(d_{ij})$. The Y_{obs} corresponding to the nearest neighbour obtained from the above equation is used as the imputed value for Y_{mis}. We term this approach as nearest neighbour by propensity score (NNPS).

5.5.2 Graphical Methods

For the covariate selection by graphical modelling (GM) we need to find the partial correlations. The steps to find partial correlation are:

1. Generate the covariance matrix from the complete cases of the sampled data.

2. Compute the inverse of the covariance matrix.

3. Compute the reciprocal of the diagonal elements of the inverse matrix computed in step-2 to get the partial variances.

4. Multiply the square root of the partial variances with the inverse matrix in step-2. This makes the diagonal elements equal to one and scales the off diagonal elements (Whittaker, 1990 p.156).

The matrix obtained from these transformations is the matrix of partial correlation coefficients. This matrix of partial correlations is a $(v+1) \times (v+1)$ symmetric matrix with diagonal elements equal to unity. Here v is the number of variables in the data. In our analysis we use the partial correlations related to the dependent variable (HL) because we are interested in which set of covariates best explains the variation in HL. The partial correlation between particular variables is used to test if any of the sample partial correlation $\hat{\rho} = 0$, where $\hat{\rho} = 0$ that the variables are conditionally independent. To test this we use

$$\text{dev} = -\text{nlog}(1 - \text{corr}(HL, X_k|\text{rest})^2) \; \forall k = 1, 2...v$$

where $\text{corr}(HL, X_k|\text{rest})$ is the observed partial correlation between the variable of interest (HL) and a particular covariate X_k conditioned on the covariates other than X_k. This test is repeated v times since we delete a single edge at a time. We use this test as recommended by Whittaker (1990, p-189). The above test has chi-squared distribution with one degree of freedom.

There are problems in applying the test for multiple testing of partial correlations as done in our work in that though each test is made at the nominal 5% significance level, the overall test has a much higher and unknown significance level (Whittaker, 1990). Another problem with the graphical modelling is the computation of partial

correlation depends on the covariance matrix and sometimes the covariance matrix may be singular.

5.5.3 Principal Components

For the data reduction by principal component analysis we used the subselect package developed by Cerdeira *et al* (2004), and available in R. This package has the advantage of retaining a variable in the subset if its importance is known a priori. For the computation of eigen values and eigen vectors, the requirement of the package was to supply a symmetric matrix (covariance /correlation). This could be a correlation or a covariance matrix. We conducted subset selection using both the covariance matrix and the correlation matrix. The generalized coefficient of determination (GCD), and the subset of variables as predictors (SVP) approaches to principal component analysis as described in chapter 4 section 4.4 were used for selecting the subsets of the data using correlation matrix. For data reduction by PCA we specified the number of covariates including the dependent variable, to be seven. This was the number of covariates chosen by graphical modelling. By keeping the same number of covariates, we make the size of the covariates in the model to be the same for both PCA and GM. This ensures that difference between the results of the two different data reduction method occur from their different methodology, not that one has more covariates to predict than the other. A list of covariates selected by graphical models and by PCA in simulations is given in table-5.4.

5.5.4 Results

The main purpose of the data reduction is to reduce the computation time for MVNN while maintaining the quality of the imputation. Table-5.5 presents the time taken

Table 5.4: Six Commonly selected covariates under Principal Component Analysis and Graphical Modelling

Method	List of covariates
PCA	
GCD	Region, CEB, SSLI,Drnkal, Current pregnant, Chickmeat.
SVP	Region, CEB, SSLI,Drnkal, Menlsw, Eggs
Graphical	Region, CEB, Current pregnant, Suff jaundice, Chew Tobacco, logalt.

Table 5.5: Computation time in seconds: Data imputed by MVNN and NNPS methods

	Nearest neighbour		
Sample size	All covariates	(GM, PCA)	NNPS
300	29.75	7.72	1.88
500	133.8	26.75	3.18

for one simulation with a sample size of 300 and 500. In this simulation we used all the covariates, subset of covariates obtained by PCA, graphical modelling and NNPS. These simulations were conducted on a Pentium III under the Windows environment. The times given in the table-5.5 are for one imputation on a single variable. For a sample size of 300 and using 28 covariates it approximately took 30 seconds to do the imputation. In large scale surveys there will be k variables to impute hence the computation time would be increased by a factor of k (The number of variables that have nonresponse). Also the sample size will be much larger (e.g. NFHS-2 is 9292 for U.P). Sometimes after the initial imputation there may be for some reasons, for example additions or deletions to the list of covariates,

so this process may be repeated several times. Keeping this in mind we investigated data reduction methods to improve the computation. Using a subset of covariates the imputation time has reduced to about a fifth with propensity matching is being considerably faster than the other subset selection methods.

Having shown there is a significant reduction in computation time using data reduction methods, we now need to look at the quality of the imputed data. For these comparisons we use the data imputed using all variables as the benchmark method, as we cannot do better as we have used all the covariates available to impute. For the first set of comparisons we used the correlation matrix for subset selection by PCA as well as graphical modelling and our newly developed propensity scores. In the tables it is referred to as nearest neighbour by propensity score (NNPS). As in our other work earlier these comparisons were carried out for various nonresponse rates (15%, and 25%) and for various sample sizes (300 and 500). For studying the performance we use the four measures used in earlier analysis (MSEM, MSEI, MAD and Leti's index).

5.5.5 Summary of the Data Reduction Methods

For our initial study of the data reduction methods for imputation we used a data sets (Low Birth Weight) from Hosmer and Lemshow (2002). In the simulations of the low birth weight we used simple MAR response mechanism to study and compare the performance of the MVNN, RBNN and NNPS methods. The results were encouraging and are reported in Murthy and Chacko (2004) and given in Appendix-E. These results motivated us to further investigate the performance of data reduction methods, especially NNPS described above, using the NFHS-2 data. In the simulations of NFHS-2 data, we use different MAR models, percentages of nonresponse and sample sizes. The results of these comparisons are presented in Tables(5.6-5.8).

In these simulations we again used MSEM, MSEI, MAD and Leti's index to study the performance of the different methods. The results in Tables and box plots are in Figures (5.12-5.21). They show very little differences in the distribution in most of the comparisons, indicating that imputing using a subset of covariates is similar to imputing using the full set of covariates. Sometimes, for example when the MAR model is simple MAR and sample sizes was 300, there are some unexpected observations. The bias is considerably lower for one or two cases. This is present for the data imputed using the subset of covariates as well as the full set of covariates under 5% and 15% nonresponse. This may be because that some unimportant variables may dominate the distance and lead to wrong match. This was noticed in simple MAR and also at times in the other three MAR mechanisms. However the reason for this is not clear form the current simulations. Further studies are planned to try and understand this behaviour.

The differences between the variable reduction methods do not appear to be statistically significant to identify an ideal subset method. Considering the reduction (of up to five times) of computational time and that there is very little difference in the performance of the reduced subset methods, it is worth considering variable reduction. Among these NNPS is fastest and often appears to perform better than the other subset selection methods. Also, as discussed in (see section 4.6), since NNPS uses propensity scores which has some good properties, we would recommend the use of NNPS for data reduction methods discussed here.

Table 5.6: Comparisons of the performance of the data imputed using all variables and subset of variables: under the simple MAR model

Sample Size 300									
Method	Nonresponse								
	5%			10%			15%		
	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD
MVNN	0.62	26.1(0.4)	20.94	0.78	25.2(0.2)	19.91	0.65	25.3(0.2)	18.69
NNPS	0.52	24.8(0.5)	20.60	0.51	25.3(0.5)	20.26	0.71	25.7(0.4)	20.38
GM	0.42	24.3(0.5)	20.06	0.54	25.1(0.4)	20.35	0.75	24.8(0.3)	20.99
PCA									
GCD	0.40	24.8(0.5)	20.38	0.52	25.7(0.4)	19.50	0.76	25.3(0.5)	20.15
SVP	0.36	25.6(0.4)	19.25	0.52	25.2(0.3)	19.85	0.76	25.8(0.4)	19.50
Sample Size 500									
MVNN	0.17	25.7(0.3)	20.08	0.39	24.4(0.1)	19.05	0.37	24.1(0.2)	19.59
NNPS	0.34	25.9(0.4)	20.60	0.51	25.2(0.4)	20.64	0.57	25.4(0.4)	20.52
GM	0.28	25.3(0.4)	20.79	0.45	25.6(0.3)	20.87	0.67	25.1(0.3)	20.83
PCA									
GCD	0.26	25.4(0.4)	20.60	0.42	25.7(0.3)	20.51	0.60	25.4(0.3)	20.52
SVP	0.28	26.3(0.4)	19.90	0.42	26.2(0.3)	20.35	0.62	26.3(0.3)	20.53

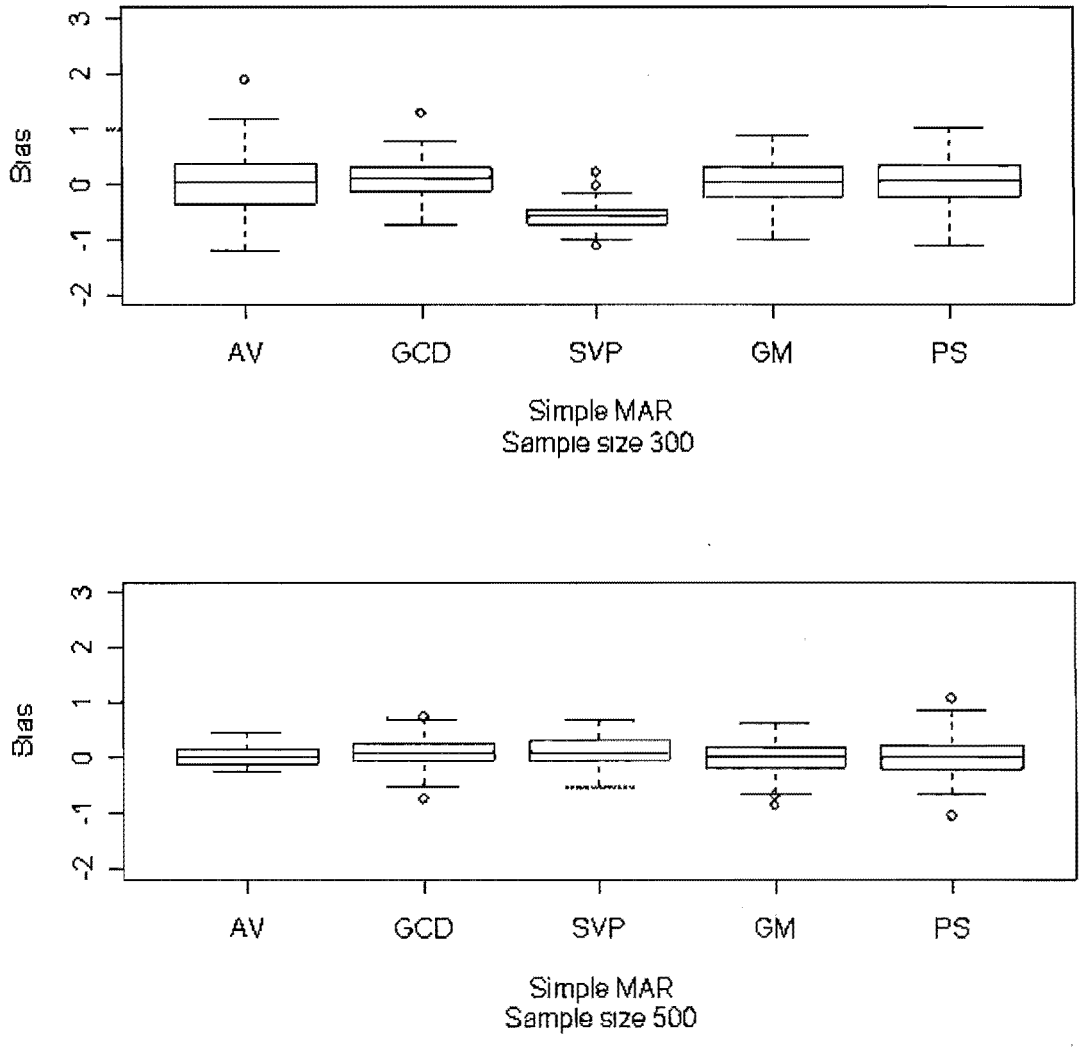

Figure 5.12: Distribution of bias for 5% nonresponse

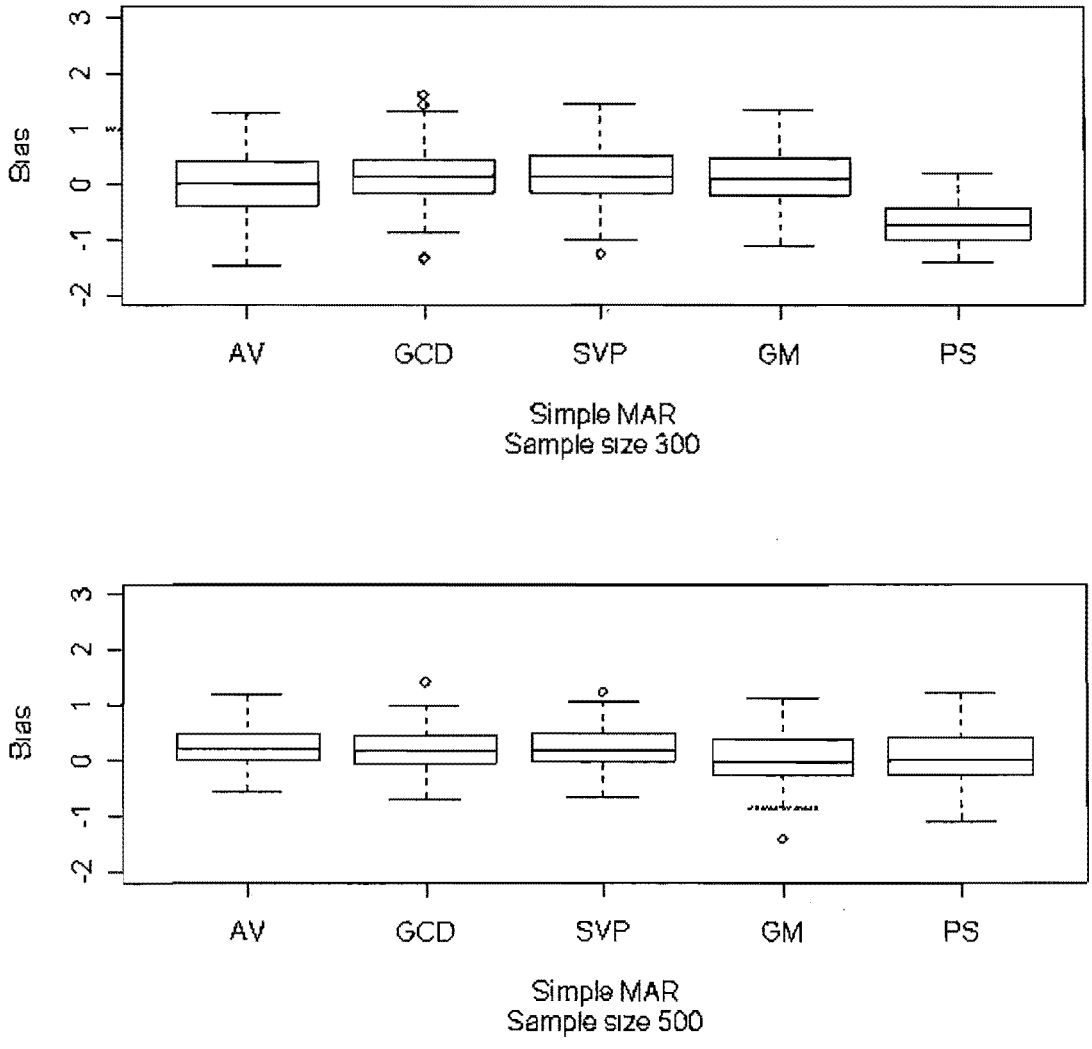

Figure 5.13: Distribution of bias for 10% nonresponse

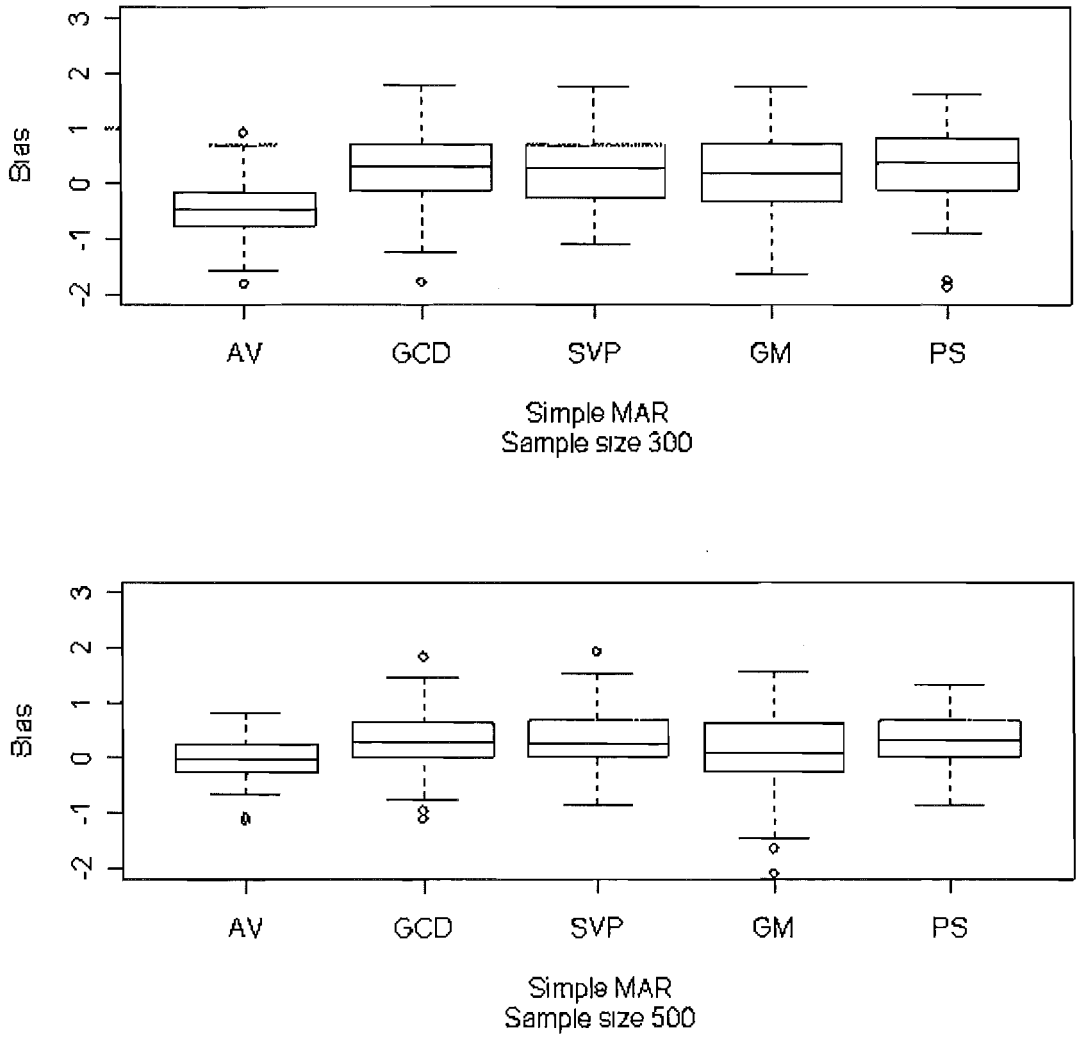

Figure 5.14: Distribution of bias for 15% nonresponse

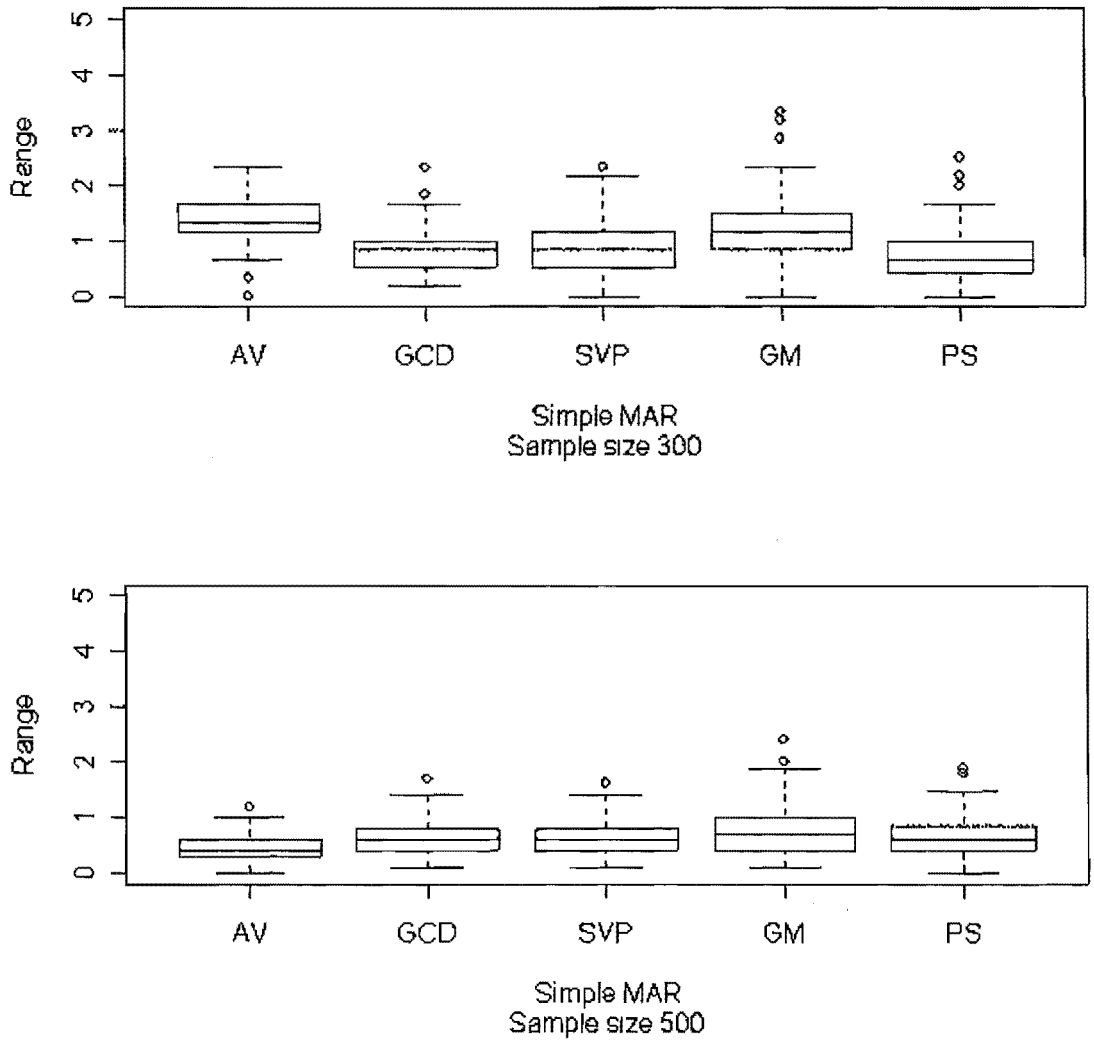

Figure 5.15: Distribution of Leti's index for 5% nonresponse

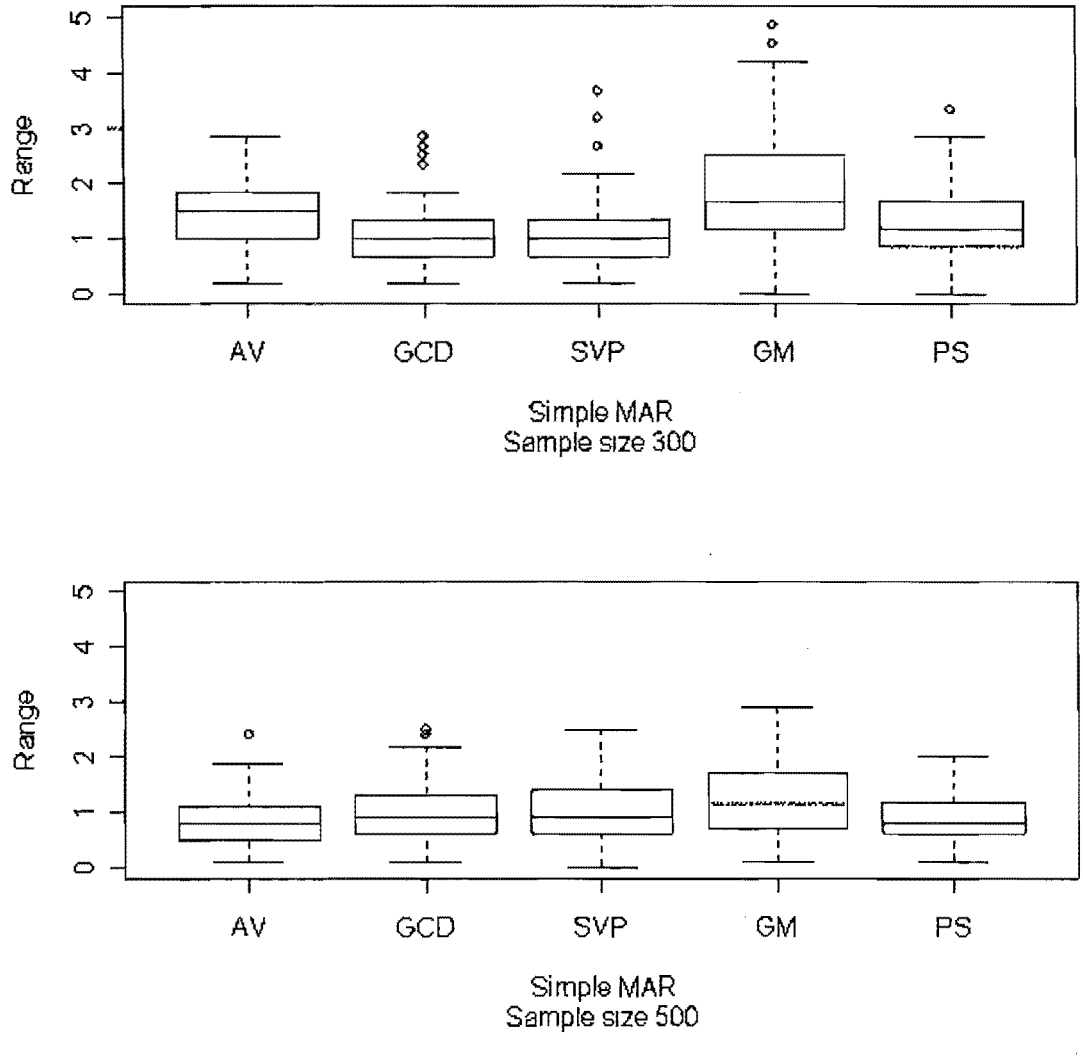

Figure 5.16: Distribution of Leti's index for 10% nonresponse

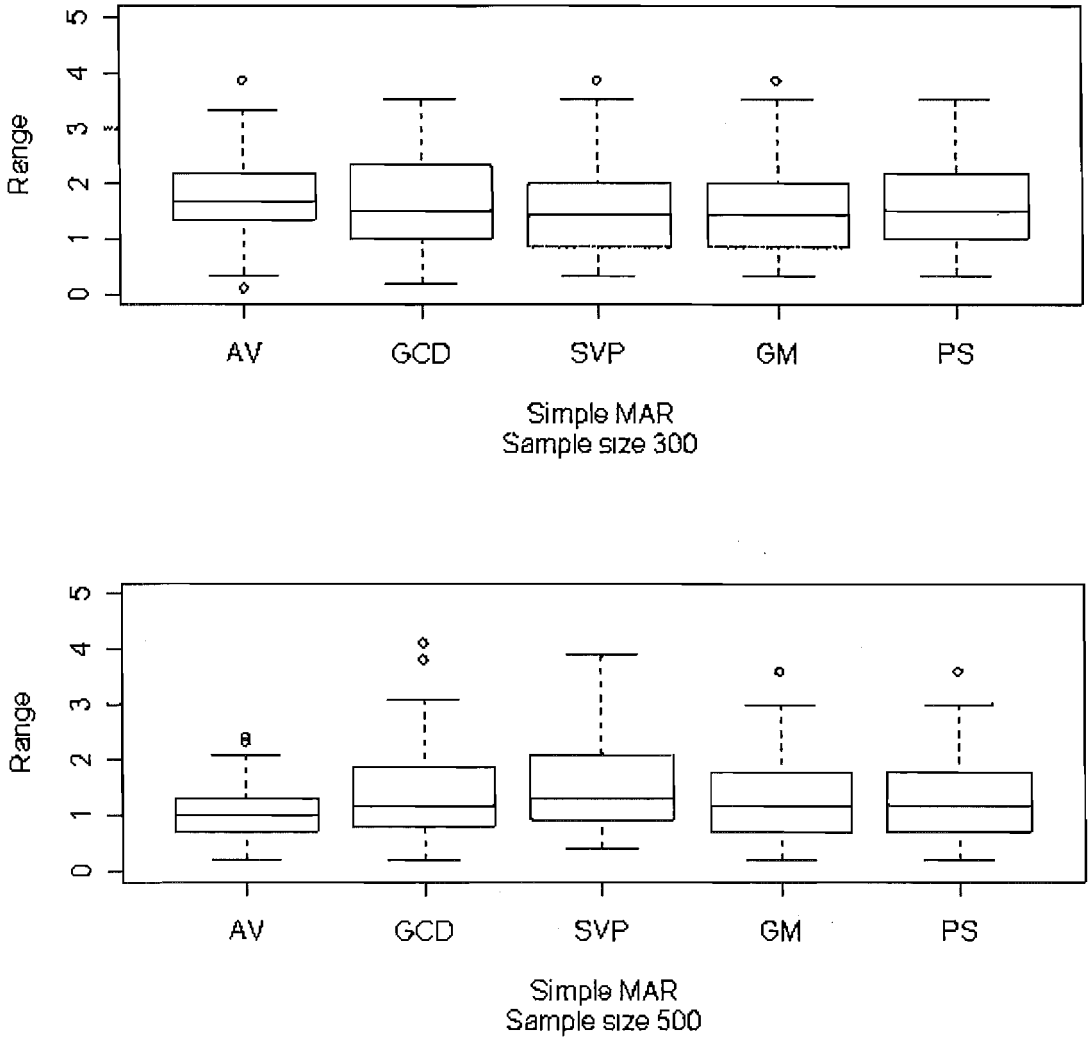

Figure 5.17: Distribution of Leti's index for 15% nonresponse

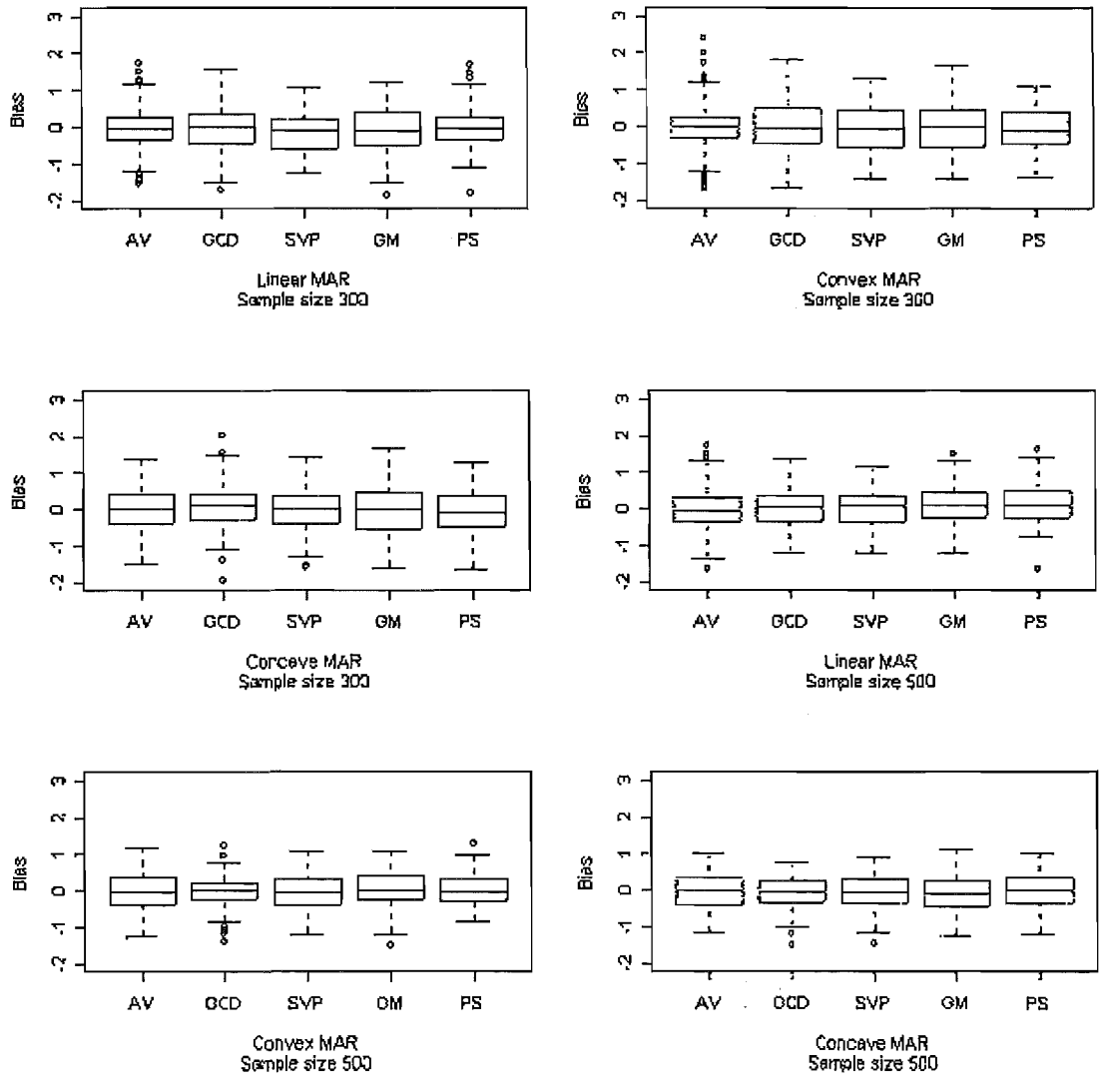

Figure 5.18: Distribution of bias for 15% nonresponse

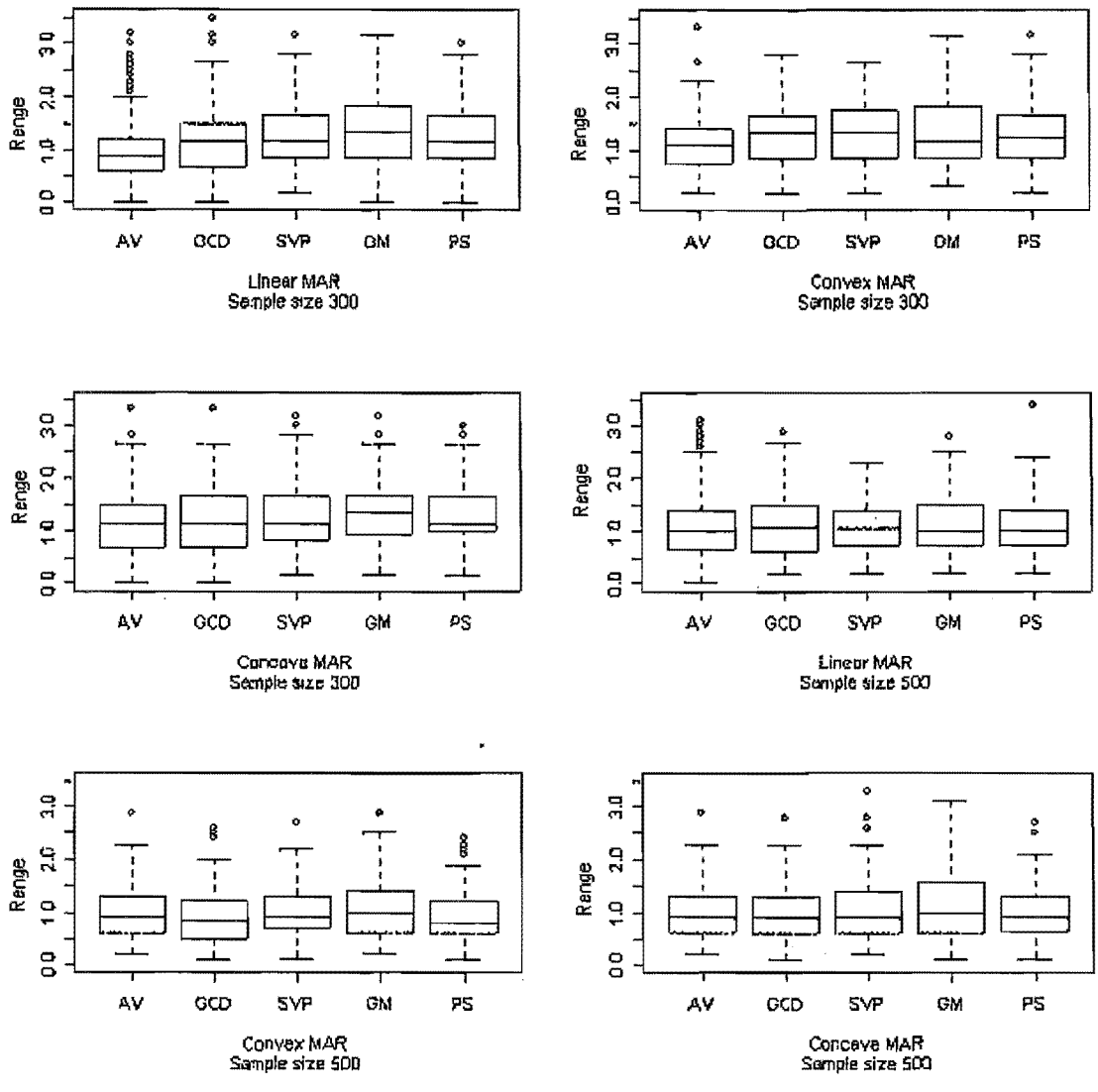

Figure 5.19: Distribution of Leti's index for 15% nonresponse

Table 5.7: Comparison of the performance of the data imputed using all variables and subset of variables: 15% nonresponse data

Sample Size 300									
	Type of missing								
Data	Linear			Convex			Concave		
Reduction	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD
AV	0.46	26.6(0.1)	20.77	0.51	25.8(0.1)	20.54	0.58	26.7(0.1)	20.72
NNPS	0.60	27.3(0.4)	21.42	0.59	27.1(0.4)	21.35	0.56	26.9(0.4)	20.20
GM	0.63	27.1(0.3)	21.35	0.67	27.2(0.5)	21.33	0.67	27.0(0.4)	20.74
PCA									
GCD	0.60	26.8(0.4)	21.08	0.67	26.6(0.4)	20.92	0.59	27.1(0.4)	21.29
SVP	0.55	26.9(0.4)	21.18	0.62	26.7(0.4)	20.99	0.58	26.7(0.4)	20.74
Sample size 500									
AV	0.51	26.5(0.1)	20.65	0.48	26.7(0.1)	20.60	0.50	27.1(0.1)	20.85
NNPS	0.55	26.2(0.2)	20.60	0.46	26.6(0.3)	20.85	0.48	26.9(0.3)	20.79
GM	0.53	26.4(0.3)	20.67	0.53	26.4(0.3)	20.80	0.54	26.6(0.3)	21.05
PCA									
GCD	0.53	26.6(0.3)	20.90	0.44	26.6(0.3)	20.80	0.43	26.8(0.3)	20.74
SVP	0.51	26.5(0.3)	20.85	0.48	26.4(0.4)	20.75	0.48	26.7(0.3)	20.65

Table 5.8: Comparison of the performance of the data imputed using all variables and subset of variables: 25% nonresponse

Sample Size 300									
	Type of missing								
Method	Linear			Convex			Concave		
	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD
AV	0.91	26.8(0.1)	20.70	0.95	26.6(0.1)	20.64	0.92	26.4(0.1)	20.65
NNPS	1.03	26.3(0.3)	20.50	0.94	27.1(0.3)	20.87	0.91	27.1(0.3)	21.24
GM	1.06	26.4(0.4)	20.75	0.85	27.1(0.5)	20.76	1.05	27.3(0.4)	21.27
PCA									
GCD	0.92	26.2(0.5)	20.84	0.80	26.6(0.6)	20.71	0.90	27.1(0.4)	21.02
SVP	0.83	26.2(0.3)	20.48	0.85	27.2(0.4)	21.08	0.92	27.2(0.5)	21.06
Sample size 500									
AV	0.67	26.8(0.1)	20.87	0.65	26.6(0.1)	20.75	0.68	26.6(0.1)	20.55
NNPS	0.71	26.7(0.2)	20.87	0.70	26.9(0.3)	20.84	0.70	26.7(0.2)	20.58
GM	0.72	27.2(0.3)	21.09	0.83	27.3(0.4)	21.13	0.84	27.2(0.3)	20.85
PCA									
GCD	0.75	27.1(0.2)	21.02	0.70	27.1(0.3)	20.86	0.73	27.3(0.3)	20.65
SVP	0.76	27.3(0.3)	21.00	0.72	27.1(0.3)	20.93	0.70	27.1(0.3)	20.63

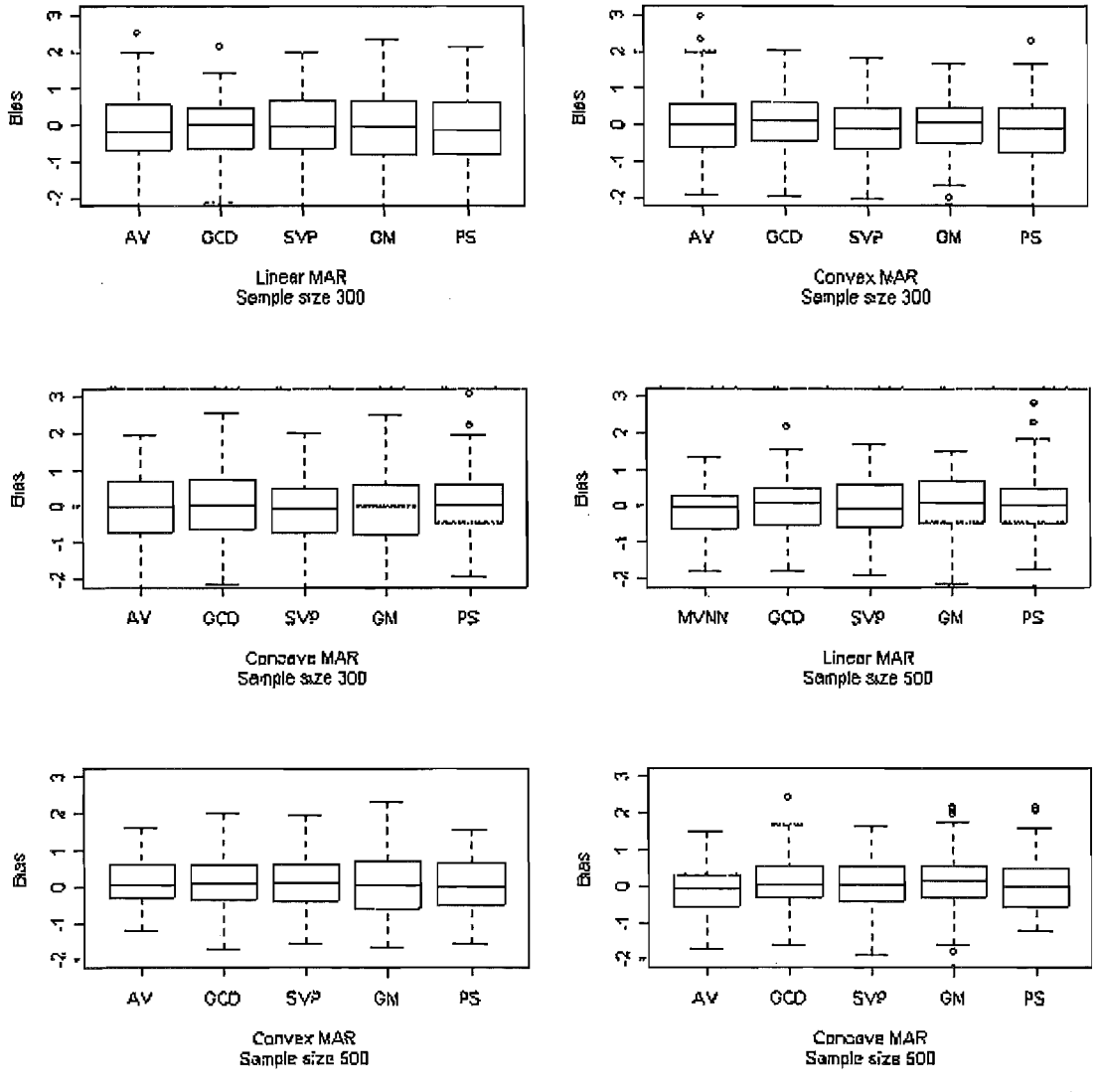

Figure 5.20: Distribution of bias for 25% nonresponse

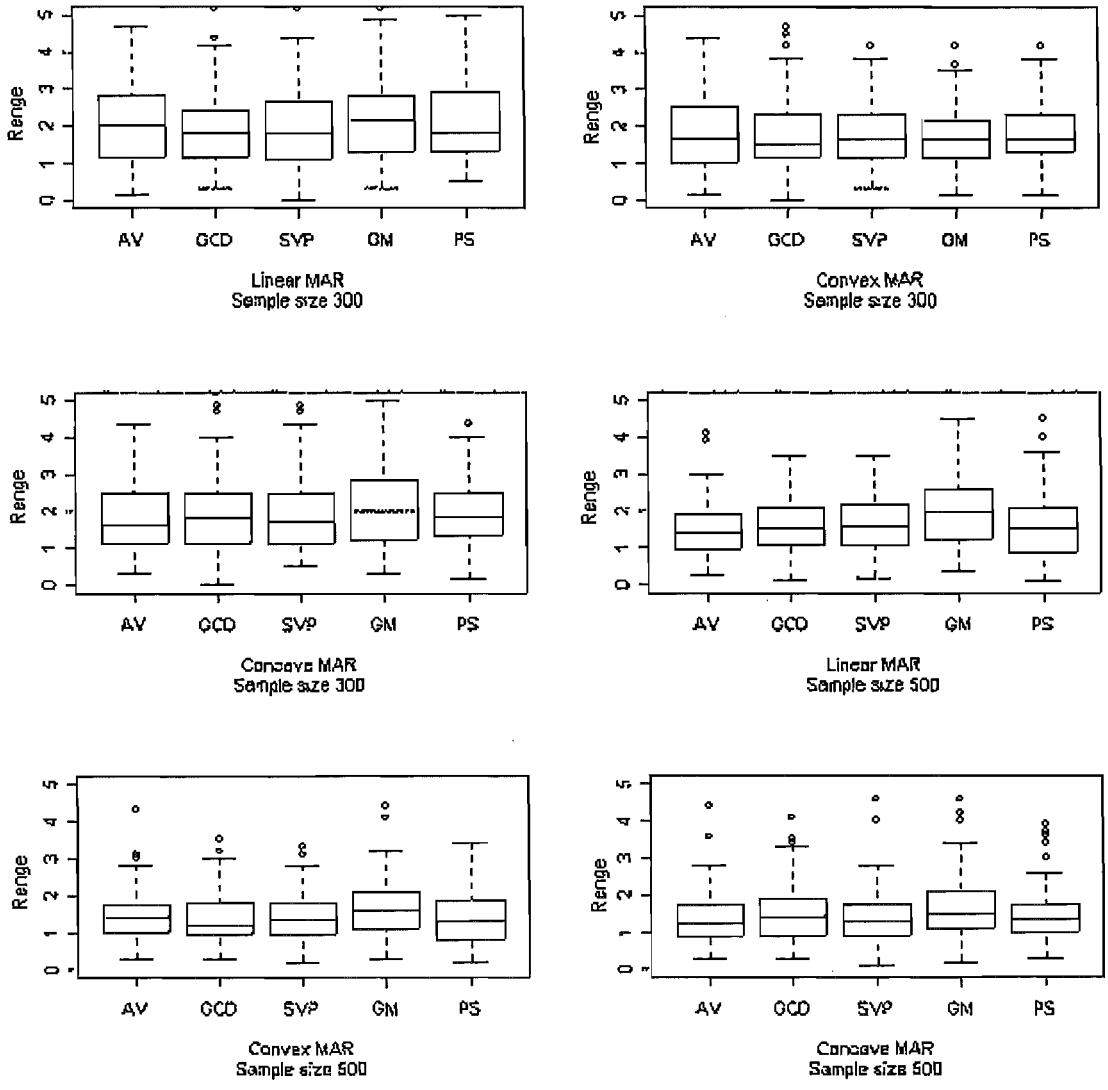

Figure 5.21: Distribution of Leti's index for 25% nonresponse

5.6 Subset by Correlation and Covariance

Generally in the multivariate analysis data reduction is done by PCA using a covariance or a correlation matrix. In NFHS-2 data we have mixture of variables measured on different scales and it is not clear whether the covariance or correlation should be used. We therefore ran some tests on the use of the correlation as well as the covariance matrix for selecting the subset of covariates. In this section we discuss the comparison of variable reduction using these two methods.

As there were no significant differences in the results obtained using either the GCD or SVP criteria in the previous analysis, we use only GCD in these comparisons. The subset obtained using the covariance matrix are Region, CEB, Age, Caste, Eggs and Education, whereas the subsets obtained using correlation matrix are Region, CEB, Current pregnancy, Chicken or meat, standard of living and drink alcohol. Note that there are only two common covariates selected using these two matrices. For imputing the missing data in the data quality we used our MVNN preferred methodology imputation. As before we study the performance using all four quality measures and the four nonresponse mechanisms. Results are presented in Tables-(5.9-5.10) for both 15% and 25% nonresponse.

We observe that while there are differences between the subsets obtained using the correlation or covariance matrix the results of the imputation are not significantly different. In the box plots of the bias, GCDC refers to the covariance matrix and GCD to the correlation matrix. From figure-5.22 and figure-5.24 we see that the distributions of bias are very similar for data imputed by the subsets obtained using both the matrices. From the box plots presented in figure-5.23 and figure-5.25 we observe that the subset data obtained using covariance matrix has higher ranges of

Table 5.9: Comparison of the performance of the data imputed using covariates selected by covariance matrix and correlation matrix in PCA using GCD criteria: 15% nonresponse

Sample Size 300									
	Type of missing								
Matrix	Linear			Convex			Concave		
	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD
Correlation	0.60	26.8(0.4)	21.08	0.67	26.6(0.4)	20.92	0.59	27.1(0.4)	21.29
Covariance	0.63	27.1(0.1)	21.38	0.59	26.8(0.2)	20.96	0.56	26.9(0.4)	20.80
Sample size 500									
Correlation	0.53	26.6(0.3)	20.90	0.44	26.6(0.3)	20.80	0.43	26.8(0.3)	20.74
Covariance	0.47	26.6(0.1)	20.78	0.50	26.5(0.1)	20.65	0.48	26.9(0.3)	20.96

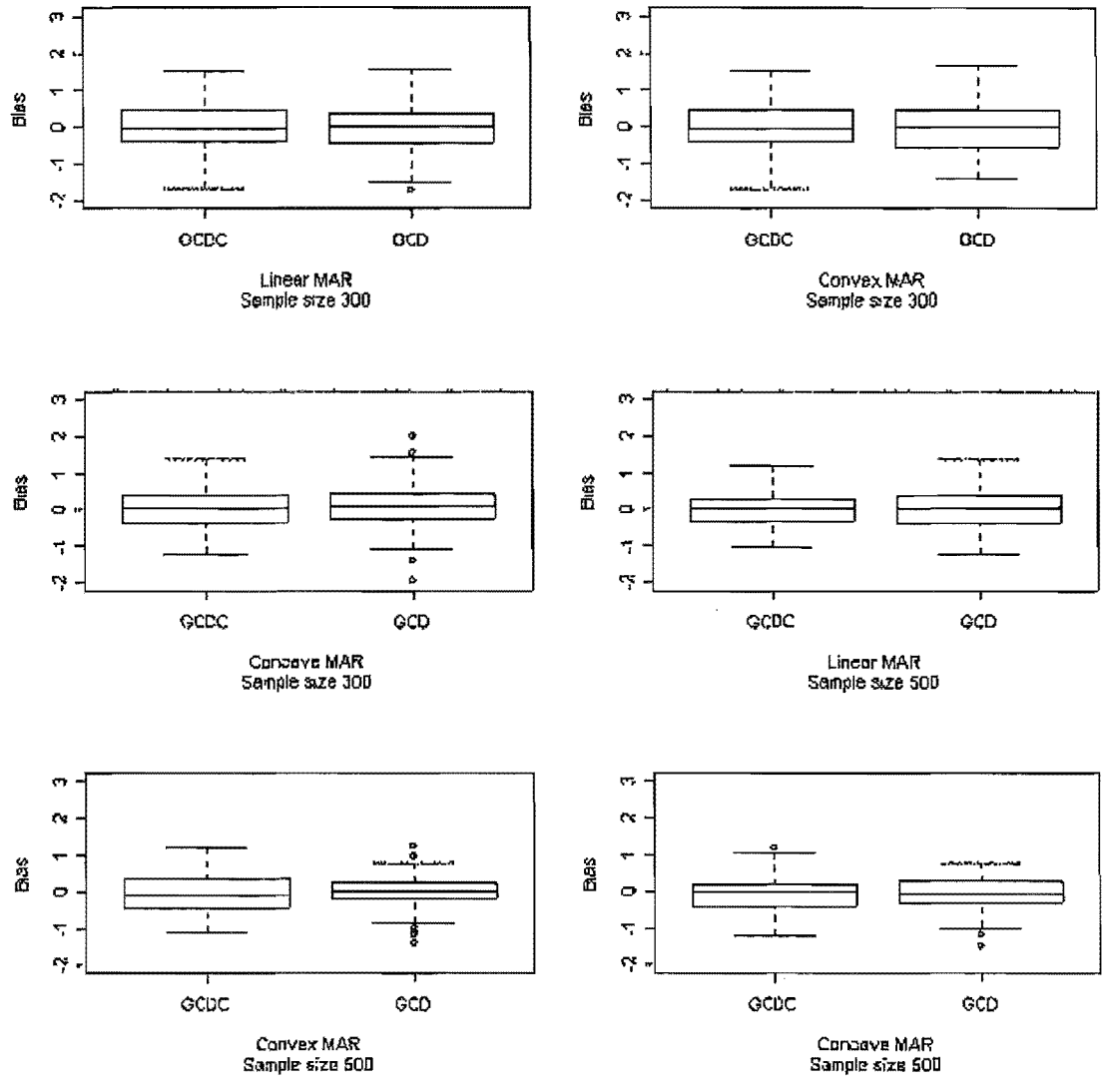

Figure 5.22: Distribution of bias for 15% nonresponse

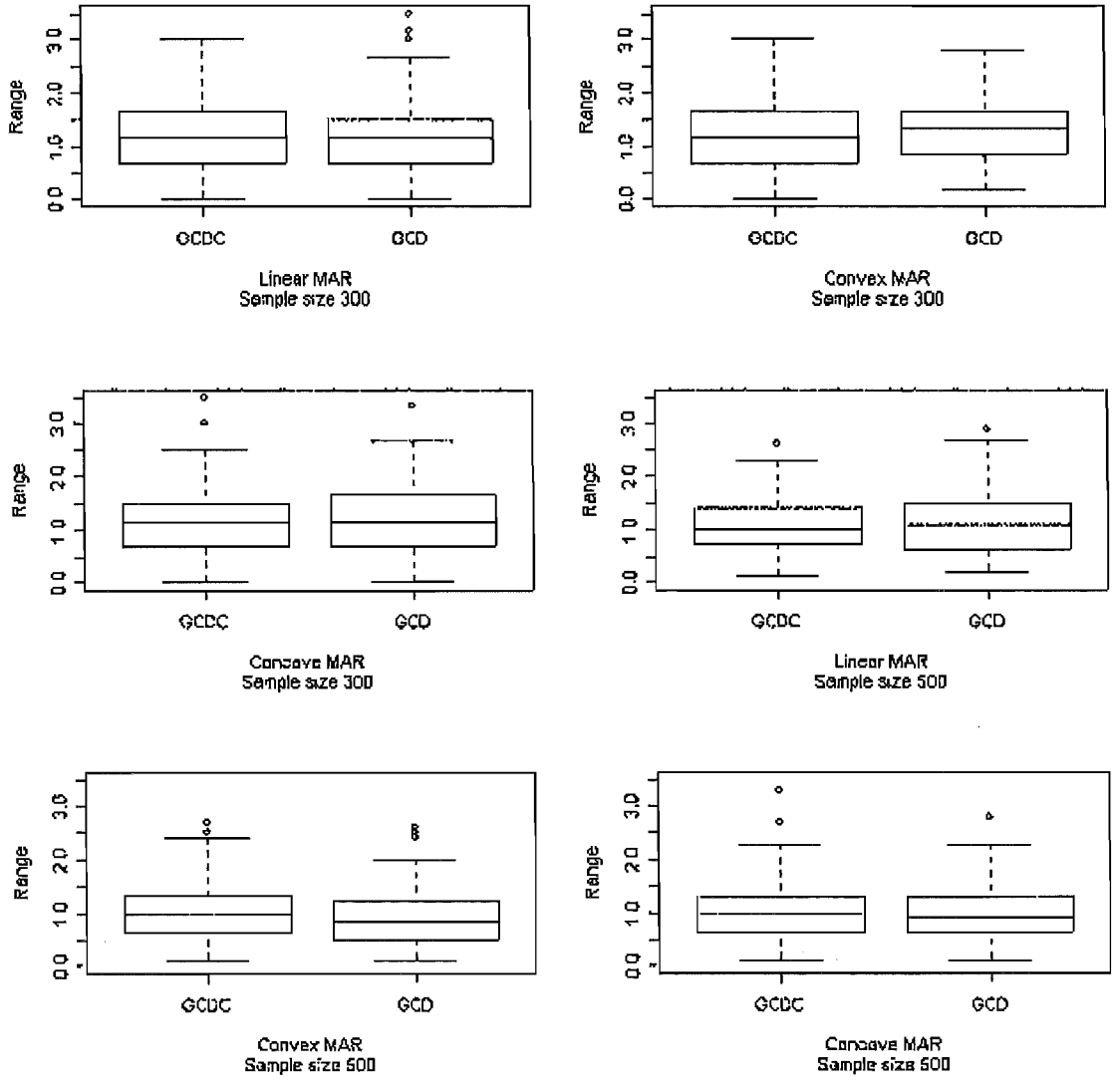

Figure 5.23: Distribution of Leti's index for 15% nonresponse

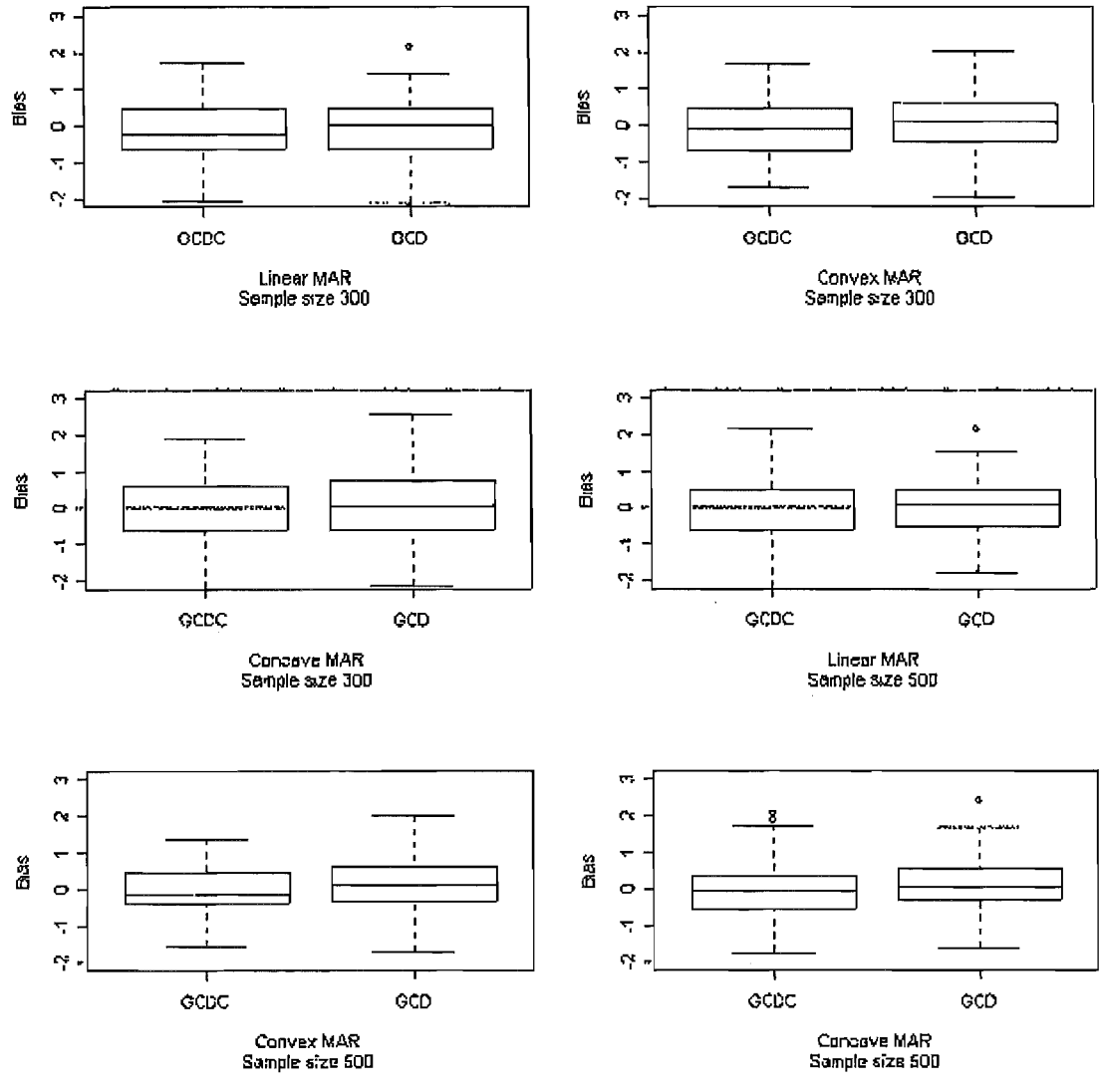

Figure 5.24: Distribution of bias for 25% nonresponse

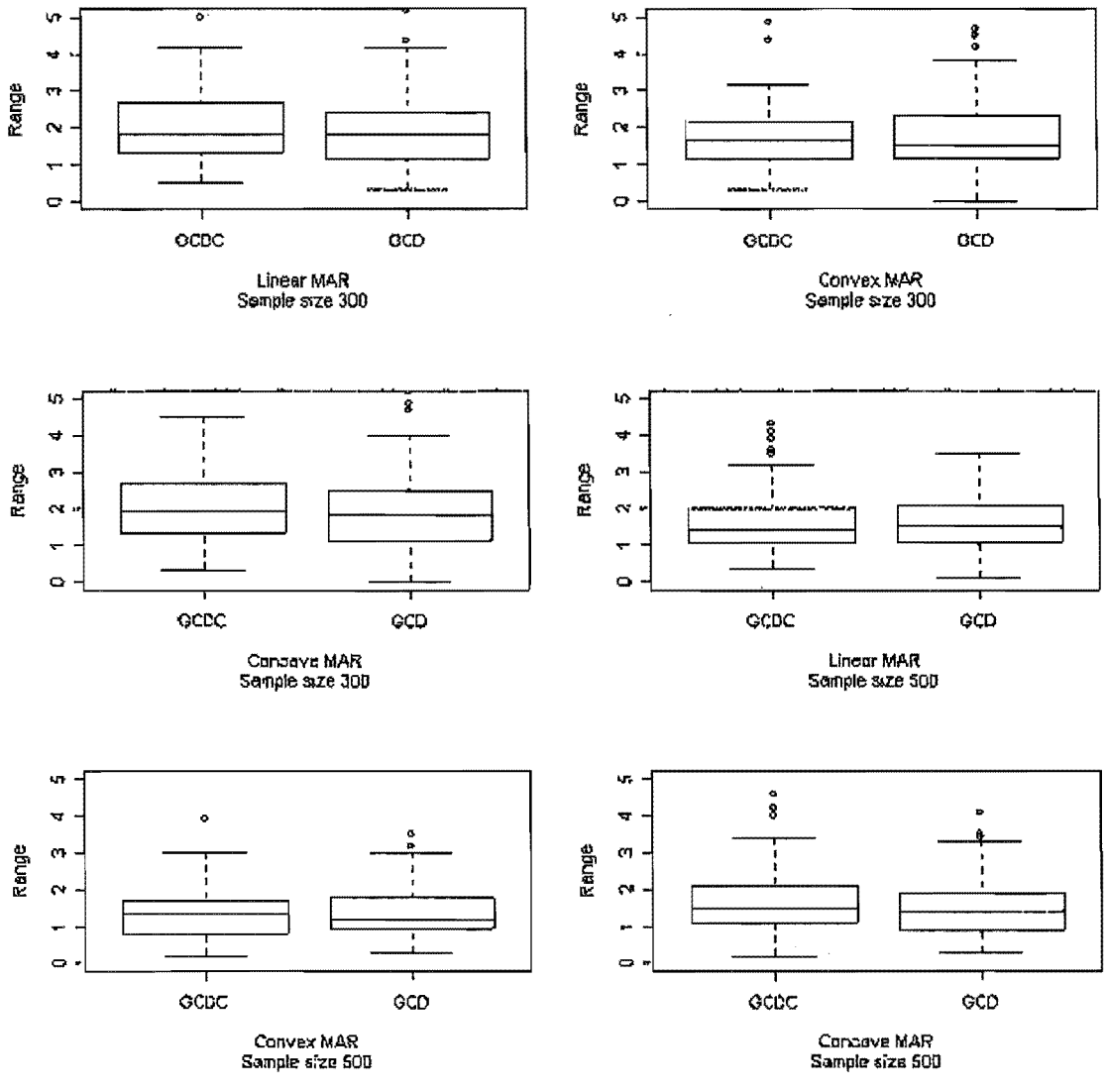

Figure 5.25: Distribution of Leti's index for 25% nonresponse

Table 5.10: Comparison of the performance of the data imputed using covariates selected by a covariance matrix and correlation matrix in PCA using GCD Criteria: 25% nonresponse

Sample Size 300									
	Type of missing								
Matrix	Linear			Convex			Concave		
	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD	MSEM	MSEI(se)	MAD
Correlation	0.92	26.2(0.5)	20.84	0.80	26.6(0.6)	20.71	0.90	27.1(0.4)	21.02
Covariance	0.92	26.4(0.2)	20.55	0.86	26.7(0.3)	20.87	0.93	26.9(0.2)	21.05
Sample size 500									
Correlation	0.75	27.1(0.2)	21.02	0.70	27.1(0.3)	20.86	0.73	27.3(0.3)	20.65
Covariance	0.79	26.8(0.1)	20.88	0.69	26.9(0.1)	21.03	0.71	26.3(0.2)	20.65

Leti's index compared to the subset obtained using correlation matrix, but as the sample size increased there is little difference.

Even though we did not find any distinguishable differences in the performance of imputation using the subset selected by covariance matrix or by correlation matrix in the PCA analysis we do not recommend the use of covariance matrix when there are mixed type of variables. In this case the lack of differences may be due to the fact that the two common variables in fact they are the two variables selected by all the methods are the only ones that are important in this case. We suspect that imputation these two variables alone as covariates may give good results but this has to be investigated further.

5.7 Discussion

In this chapter we compared the performance of our MVNN methodology and NNPS using different sample sizes, different response rates. It should be noted that using NNPS also we find the nearest neighbour, since it is only one variable it is like a subset of the MVNN computation. MAR simple is often the only missing data mechanism used in most published work. Collins *et al* (2002) appear to be the first to study varying types of MAR models in their study of multiple imputation methods. We have extended this idea to test the performance of our imputation methods using different MAR response models.

Under simple MAR we observe that MVNN performs better than the RBNN method. For other MAR-mechanisms the differences are not significant in any one comparison but generally MVNN performs better. MVNN has a further advantage in that it is nonparametric and thus avoids model misspecification errors and handles mixed types of covariates easily. Thus we recommend MVNN as the preferred method for imputation.

However MVNN is computationally intensive, hence we looked at subset selection using graphical modelling, principal component analysis and propensity scores. Using the subset of the covariate has brought a considerable reduction in computation time without distorting the quality of the results of the MVNN method.

Note from Table-5.4 that the variables Region and CEB which by the way we simulated the nonresponse in the data are good explanatory covariates of the data, as they are always selected. Also variables such as caste, nutrition (eggs, chickmeat) selected by one or other of the methods are described in medical literature as having an influence on HL. Thus both guideline i) and ii) of Sixten and Sarndal (2002) are satisfied. This may explain why we observed no significant differences when the subsets were used.

From these results we conclude that:

Although the literature on missing data has emphasized evaluating missing data in terms of bias, it useful not only to examine the bias but also to study the general closeness of the imputed values to the true values and whether the marginal distributions are preserved. This is because these measures may well be relevant to the outcomes of the preserved. This is because these measures may well be relevant to the out comes of the analysis performed on the impute data.

From the data reduction work we have done it seems that Region and CEB are the two important variables that have influence on HL. This leads us to believe more than one data reduction should be used to better see which covariates are useful predictors of imputed values.

Generally MVNN and RBNN have similar performance. However, MVNN avoids model misspecification and handles both categorical and continuous variable easily.

For small data sets and high nonresponse one must be careful in using MVNN or RBNN. Here data reduction may be advantageous.

For small sample sizes and low nonresponse levels and for the various MAR models, both MVNN and RBNN performs similarly but MVNN has other advantages as mentioned above.

For large samples MVNN has a smaller bias than RBNN across all MAR mechanisms. But if computation time is a problem, a data reduction method, especially NNPS, is recommended.

The chart in figure 5.26 summarizes our recommendations for choice of an imputation method.

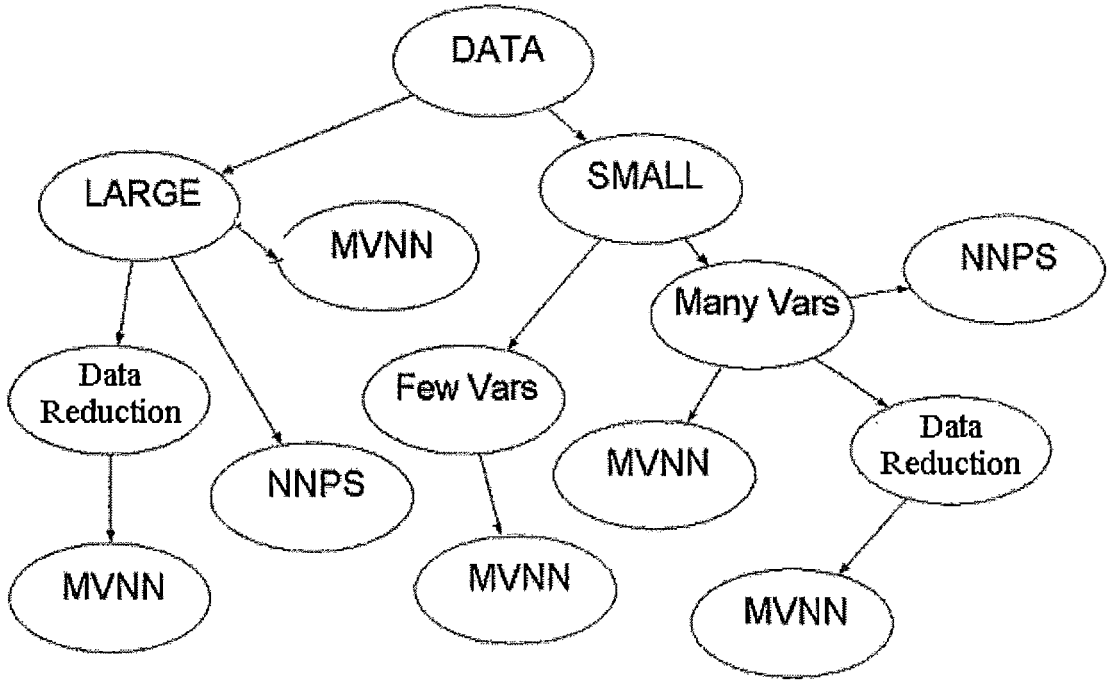

Figure 5.26: Summary chart

Chapter 6

Variance estimation

6.1 Introduction

In previous chapters we described the methods that are used for imputing the missing values and thereby creating complete data sets. Results obtained from the application of our new method (MVNN) are presented in chapter 5. Many analysts treat imputed data as real data in their analysis. Making inferences on such a basis leads to underestimation of the true variance of the estimators derived from the imputed data, since additional variability due to the unknown missing values being replaced with imputed values is not taken into account (Rao, and Shao, 1992; Chen and Shao, 2001; Rancourt *et al*, 1994) .

This additional variability due to imputation has usually been studied using three approaches:

1. Design-based approach. Under this approach the variance under consideration is with respect to a design which is used for sample selection from a fixed finite population, with necessary adjustments for additional variance due to imputation. Under this approach resampling methods such as jackknife and

bootstrap are commonly used (Rao and Shao, 1992; Rao and Sitter, 1995; Rao, 1993; Chen and Shao, 2000; Shao and Sitter, 1996; Burns, 1990) .

2. Model assisted approach. That is, estimating variance under the consideration of a design used in repeated sampling and a model that generates the finite population and nonrespondents. For data imputed using stochastic imputation this is another alternative for estimating the variance (Sarndal, 1990; Deville and Sarndal,1992; Fay 1994).

3. Variance estimation under a Bayesian approach. Under this approach the imputations are repeated several times, using these several sets of imputed data variance is estimated. Rubin (1976) advocated multiple imputation under Bayesian approach.

Model assisted approaches are more sensitive to deviations from the model assumptions when the imputation model is implicit. Since we do not use multiple imputation for imputing the missing data, Bayesian approach does not seems to be an appropriate way of estimating the variance. Hence in this thesis, we use resampling methods for finding the variance of the singly imputed data using a stochastic imputation method. In resampling we use bootstrap method for estimating the variance. The reasons for using this method are described in later sections of this chapter.

The aim of this chapter is to find an estimator for the variance of data imputed using nearest neighbour imputation. Section 6.2 presents the details of the sample response path that is usually assumed when estimating the variance using either of the above three approaches. In section 6.3 we present a bootstrap method for estimating the variance when we have full response as well as the method of Shao and Sitter(1996) for the situation where the data has imputed values. Shao and Sitter's

method can have problems when the data sets are small and the nonresponse is high or when the data sets are large and the nonresponse is low. To overcome this problem we suggest a modification to Shao and Sitter's method. This modified bootstrap which we propose is presented in section 6.4. A possible extension of our bootstrap method for the case of cluster sample designs and this outlined in section 6.5. Results and conclusions are presented in section 6.6 where we show that SS and the modification are suitable variance estimation methods.

6.2 Overview of Variance Estimation Methods

As stated in the introduction, variance methods defined on the assumption of full response, if used for estimating the variance of the data which contains imputed values will give the wrong estimates of variance. This is because they do not reflect the additional variation due to using imputed values for unobserved values. The approaches proposed for estimating the total variance for data with imputed responses follow either of two paths to go from the population to the sample with missing values.

$$\begin{array}{ccc} \text{Population} & \xrightarrow{\mathcal{S}} & \text{Complete Sample} \\ \downarrow \mathcal{R} & & \downarrow \mathcal{R} \\ \text{Census with missing data} & \xrightarrow{\mathcal{S}} & \text{Sample with missing data} \end{array}$$

where $\mathcal{S}$ is the known sampling design and $\mathcal{R}$ is the, unknown, response mechanism (RM) Fay (1991).

$\mathcal{SR}$-path: in this path it is assumed that we have a complete sample from the population, and then nonresponse occurs in the selected sample.

$\mathcal{RS}$-path: in this path it is assumed that we have a Census of the population with both respondents and nonrespondents. From this census a sample is drawn and

thus the sample contains both respondents and nonrespondents.

Fay (1991) describes the $\mathcal{SR}$-path as multiple imputation inference and the $\mathcal{RS}$-path as design-based inference. For both the paths there are two components of uncertainty; that due to sampling and that due to response mechanism.

For the $\mathcal{SR}$-path we impute the missing data thus creating a complete set of data thus creating uncertainty due to imputation. For example: an imputation model could be biased, plus if it has a stochastic component then the model used for imputing missing values when repeated gives different values creating variability, thus reflecting the uncertainty due to imputation. We then treat the imputed data as complete data and use the standard complete sample variance estimation formulas to make inferences to the population and the inferences will have uncertainty due to sampling. Under this setup, generally, the sampling design is considered to be ignorable (sec.2.3.4).

For $\mathcal{RS}$-path, we estimate the census from the sample with missing data creating uncertainty due to sampling. Then we make inference from the census to the population. The census is treated as a realization from the superpopulation which then creates uncertainty in estimating superpopulation. This uncertainty is due to response mechanism. Methods for variance estimation under the $\mathcal{RS}$-path are given by Shao and Steel (1999) and Tollefson and Fuller (1992).

In the presence of nonresponse, most researchers use the $\mathcal{SR}$-path to develop methods to estimate the variance of the imputed data (Lee *et al*, 2002). The $\mathcal{SR}$-path is more intuitive since for most surveyors and analysts, nonresponse occurs after the sample s is selected (Dalenius, 1983). In this thesis also, we follow $\mathcal{SR}$-path and describe the variance estimation approaches. Under this set-up, the response mechanism (RM) is defined as a conditional probability distribution given the sample(s), and denote as $\mathcal{R}(r|s)$.

The response mechanism plays a crucial role in the theory of imputation. Point estimates (such as means and totals) can be biased if imputation is carried out under an incorrectly specified RM. The response mechanism can be MNAR also but imputation with MNAR is complex and effectively untestable. Mostly imputation methods assume missing at random (MAR) mechanism. When there is nonresponse in the data and the response mechanism is not MCAR and no adjustment is made to account for nonresponse then the parameter of interest (θ) obtained from such data would be biased. If imputation is preferred method for dealing with the nonresponse, then the imputation method which makes use of the covariates should at least reduce the bias, but is unlikely to eliminate it completely. When the imputation model estimates the nonresponse values close to the true values and the bias would be very minimal. Since the bias could be nonzero instead of finding the variance of the parameter of interest, the mean square error (MSE), which is the sum of the variance and the bias should be computed. The MSE is a more relevant indicator of the quality of the parameter obtained from imputed data ($\hat{\theta}_I$) than simply the variance (Lee *et al*, 2002).

As stated earlier we use the $\mathcal{SR}$ path approach to estimate the MSE. When evaluating expectations and variances with respect to the two stages in the $\mathcal{SR}$ path conditioning on both the stage is required

$$E_{\mathcal{SR}}(.) = E_{\mathcal{S}}[E_{\mathcal{R}}(.|s)] \tag{6.1}$$

where $E_{\mathcal{R}}(.|s)$ denotes conditional expectation with respect to the response mechanism $\mathcal{R}$ for a given sample s $E_{\mathcal{S}}(.)$ denotes expectation with respect to the sampling design $\mathcal{S}$ and $E_{\mathcal{SR}}(.)$ denotes the expectation with respect to the sampling design and response mechanism jointly (Cassel, Sarndal and Wretman, 1983). Similarly for

the joint variance, the $\mathcal{SR}$-variance or the total variance, we have

$$V_{\mathcal{SR}}(.) = E_{\mathcal{S}}[V_{\mathcal{R}}(.|s)] + V_{\mathcal{S}}[E_{\mathcal{R}}(.|s)] \tag{6.2}$$

where $V_{\mathcal{R}}(.|s)$ denotes conditional variance with respect to the response mechanism given s; $V_{\mathcal{S}}(.)$ denotes variance with respect to the sample design $\mathcal{S}$; and $V_{\mathcal{SR}}(.)$ denotes the joint variance with respect to the sample design and the response mechanism. For an estimator $\hat{\theta}_I$ of θ the estimation error is the difference $\hat{\theta}_I - \theta$. This is a random variable whose probability distribution is determined jointly by the sampling design and the response mechanism. The mean square error of $\hat{\theta}_I$ as an estimator of θ is

$$MSE_{\mathcal{SR}}(\hat{\theta}_I) = E_{\mathcal{SR}}[(\hat{\theta}_I - \theta)^2] \tag{6.3}$$

The mean square error of $\hat{\theta}_I$ can be written as the sum of the total variance and the bias squared.

$$MSE_{\mathcal{SR}}(\hat{\theta}_I) = V_{\mathcal{SR}}(\hat{\theta}_I) + [B_{\mathcal{SR}}(\hat{\theta}_I)]^2 \tag{6.4}$$

The bias is

$$B_{\mathcal{SR}}(\hat{\theta}_I) = E_{\mathcal{SR}}(\hat{\theta}_I) - \theta \tag{6.5}$$

Here $E_{\mathcal{SR}}(\hat{\theta}_I) = E_{\mathcal{S}}[E_{\mathcal{R}}(\hat{\theta}_I|s)]$ The $B_{\mathcal{SR}}(\hat{\theta}_I)$ is the bias due to sampling and the response mechanism. This arises when the parameters obtained from the imputed response data do not agree to the population parameters and if the sampling model is incorrect. However most sample designers work hard to ensure that the sampling bias is very small if not zero.

The total variance is

$$V_{\mathcal{SR}}(\hat{\theta}_I) = E_{\mathcal{SR}}\left[(\hat{\theta}_I - E_{\mathcal{SR}}(\hat{\theta}_I))^2\right]$$

Using (6.2) the variance term becomes

$$V_{\mathcal{SR}}(\hat{\theta}_I) = E_{\mathcal{S}}[V_{\mathcal{R}}(\hat{\theta}_I|s)] + V_{\mathcal{S}}[E_{\mathcal{R}}(\hat{\theta}_I|s)] \tag{6.6}$$

Using (6.6) and (6.5), (6.4) is

$$\begin{aligned} MSE_{\mathcal{SR}}(\hat{\theta}_I) &= E_{\mathcal{S}}[V_{\mathcal{R}}(\hat{\theta}_I|s)] + V_{\mathcal{S}}[E_{\mathcal{R}}(\hat{\theta}_I|s)] + (E_{\mathcal{SR}}(\hat{\theta}_I) - \theta)^2 && (6.7)\\ &= E_{\mathcal{S}}[V_{\mathcal{R}}(\hat{\theta}_I|s)] + V_{\mathcal{S}}[E_{\mathcal{R}}(\hat{\theta}_I|s) + (\hat{\theta}_s - \hat{\theta}_s)] + (E_{\mathcal{SR}}(\hat{\theta}_I) - \theta)^2 && (6.8)\\ &= E_{\mathcal{S}}[V_{\mathcal{R}}(\hat{\theta}_I|s)] + V_{\mathcal{S}}[E_{\mathcal{R}}(\hat{\theta}_I - \hat{\theta}_s|s) + \hat{\theta}_s] + (E_{\mathcal{SR}}(\hat{\theta}_I) - \theta)^2 && (6.9) \end{aligned}$$

As stated, in the $\mathcal{SR}$ path we first make inference from the sample with missing data to the complete data and then to the population. Given a sample s, the bias of estimate of the parameter form imputed data

$$B_{mis|s} = E_{\mathcal{R}}(\hat{\theta}_I|s) - \hat{\theta}_s$$

We now have $MSE_{\mathcal{SR}}$ as

$$MSE_{\mathcal{SR}}(\hat{\theta}_I) = E_{\mathcal{S}}[V_{\mathcal{R}}(\hat{\theta}_I|s)] + V_{\mathcal{S}}[B_{mis|s} + \hat{\theta}_s] + (E_{\mathcal{SR}}(\hat{\theta}_I) - \theta)^2 \qquad (6.10)$$

Now

$$V_{\mathcal{S}}[B_{mis|s} + \hat{\theta}_s] = V_{\mathcal{S}}(B_{mis|s}) + V_{\mathcal{S}}(\hat{\theta}_s) + 2cov_{\mathcal{S}}(\hat{\theta}_s, B_{mis|s})$$

Using this in (6.10) we have

$$MSE_{\mathcal{SR}}(\hat{\theta}_I) = E_{\mathcal{S}}[V_{\mathcal{R}}(\hat{\theta}_I|s)] + V_{\mathcal{S}}(B_{mis|s}) + V_{\mathcal{S}}(\hat{\theta}_s) + 2cov_{\mathcal{S}}(\hat{\theta}_s, B_{mis|s}) + (E_{\mathcal{SR}}(\hat{\theta}_I) - \theta)^2 \qquad (6.11)$$

Now

$$\begin{aligned} V_{\mathcal{S}}(B_{mis|s}) &= E_{\mathcal{S}}(B^2_{mis|s}) - (E_{\mathcal{S}}(B_{mis|s}))^2 && (6.12)\\ &= E_{\mathcal{S}}(B^2_{mis|s}) - \Big(E_{\mathcal{S}}(E_{\mathcal{R}}(\hat{\theta}_I - \hat{\theta}_s|s))\Big)^2 && (6.13)\\ &= E_{\mathcal{S}}(B^2_{mis|s}) - (E_{\mathcal{S}}\left(E_{\mathcal{R}}(\hat{\theta}_I|s) - \hat{\theta}_s\right))^2 && (6.14)\\ &= E_{\mathcal{S}}(B^2_{mis|s}) - (E_{\mathcal{S}}(E_{\mathcal{R}}(\hat{\theta}_I|s)) - E_{\mathcal{S}}(\hat{\theta}_s))^2 && (6.15)\\ &= E_{\mathcal{S}}(B^2_{mis|s}) - (E_{\mathcal{S}}(E_{\mathcal{R}}(\hat{\theta}_I|s)) - \theta)^2 && (6.16) \end{aligned}$$

where in going from (6.15) to (6.16) we assume $E_{\mathcal{S}}(\hat{\theta}_s) = \theta$ i.e for probability designs with no nonresponse the bias is zero. Using this in (6.11)

$$\begin{aligned} MSE_{\mathcal{SR}}(\hat{\theta}_I) = E_{\mathcal{S}}[V_{\mathcal{R}}(\hat{\theta}_I|s)] + E_{\mathcal{S}}(B^2_{mis|s}) - (E_{\mathcal{S}}(E_{\mathcal{R}}(\hat{\theta}_I|s)) - \theta)^2 \\ + V_{\mathcal{S}}(\hat{\theta}_s) + 2cov_{\mathcal{S}}(\hat{\theta}_s, B_{mis|s}) + (E_{\mathcal{SR}}(\hat{\theta}_I) - \theta)^2 \end{aligned} \tag{6.17}$$

As $E_{\mathcal{SR}}(\hat{\theta}_I) = E_{\mathcal{S}}[E_{\mathcal{R}}(\hat{\theta}_I|s)]$, the third and the fifth term cancel to give

$$MSE_{\mathcal{SR}}(\hat{\theta}_I) = E_{\mathcal{S}}[V_{\mathcal{R}}(\hat{\theta}_I|s)] + E_{\mathcal{S}}(B^2_{mis|s}) + V_{\mathcal{S}}(\hat{\theta}_s) + 2cov_{\mathcal{S}}(\hat{\theta}_s, B_{mis|s}) \tag{6.18}$$

Denote $E_{\mathcal{S}}[V_{\mathcal{R}}(\hat{\theta}_I|s)]$ by V_{imp} and $V_{\mathcal{S}}(\hat{\theta}_s)$ by V_{sam} then the MSE is

$$MSE_{\mathcal{SR}}(\hat{\theta}_I) = V_{imp} + E_{\mathcal{S}}(B^2_{mis|s}) + V_{sam} + 2cov_{\mathcal{S}}(\hat{\theta}_s, B_{mis|s}) \tag{6.19}$$

Omitting respondents from analysis will have bias when the response mechanism is not MCAR, as nonrespondents are different from respondents (as they have not responded). That is, whenever there is nonresponse, some bias is introduced if no adjustment for nonresponse is made. However imputation should minimize the bias due to nonresponse, though this will depend on how good the imputation model is. A good imputation model will use covariate information and if properly chosen will minimize $B_{mis|s}$ Chen and Shao (2000). Assuming this the second and fourth term in (6.19) can be ignored.

Thus

$$MSE_{\mathcal{SR}}(\hat{\theta}_I) = V_{imp} + V_{sam} \tag{6.20}$$

As MSE now consists only of variance components we denote this by V_{tot}, so that

$$V_{tot} = V_{imp} + V_{sam} \tag{6.21}$$

as given (without proof) in Lee *et al* (2002).

For estimating V_{tot} two commonly used methods are, model-assisted and resampling methods. We now present an overview of variance estimation under these two methods.

6.2.1 Model-assisted Approach

Sarndal (1990) first formulated the model-assisted approach. It has been used by Deville and Sarndal (1992) for finding the variance estimator for the data imputed under regression imputation. Later Rancourt *et al* (1994) used this approach for nearest neighbour imputed data, whereas Gagnon *et al* (1996) estimated the variance of the data imputed using a generalized regression estimator.
Under the model-assisted approach the probabilistic set-up observed is:

- The sample design ($\mathcal{S}$);
- the response mechanism $\mathcal{R}(r|s)$;
- the imputation model (m), which can be an explicit or an implicit model.

With the introduction of the imputation model the $\mathcal{SR}$ path is changed to $m\mathcal{SR}$ thus the variance of the estimator $\hat{\theta}_I$ in (6.21) becomes

$$E_m(V_{tot}) = E_m(V_{sam} + V_{imp}) = E_m(V_{sam}) + E_m(V_{imp}) \tag{6.22}$$

We again assume that the bias is zero as described in the earlier section and that the response mechanism is MAR.

Commonly a three phase approach is used in model assisted variance estimation. The sample design is the first phase, the second phase is the response mechanism and the third phase the imputation model. The objective of this approach is to find variance components $\hat{V}_{sam}$ and $\hat{V}_{imp}$ such that they are model unbiased estimators

for every fixed sample. Here $\hat{V}_{sam}$ is obtained from the imputed data. If the imputation method is deterministic then Sarndal suggests some adjustments for the $\hat{V}_{sam}$ (Sarndal, 1990).

6.2.2 Resampling Methods

In resampling we draw subsamples from the sample and use the distribution of the estimates from the subsamples to estimate the variance of the parameter of interest. Resampling methods have the advantage of not requiring any separate theoretical derivations for estimating the variance. These methods are computer intensive but became popular after the advent of fast computers. Resampling methods are often used for variance estimation of estimators in complex designs like multistage cluster designs assuming full response. Resampling methods include jackknife, bootstrap and the balanced repeated replicates (BRR). For jackknife resampling to compute variance we delete one case at a time from the survey sample and then compute the parameter. This process is repeated for all cases in the survey sample till we have n parameters estimated from the sample. Once these are estimated the variance is estimated using these n jackknife estimates. In Bootstrap, we draw a sample with replacement from the survey sample and compute the parameter using these samples.

However their application to the imputed data without any allowance for imputation variance would underestimate the total variance of the parameter of interest. Two resampling approaches to correct this underestimation of variance are reimputation and adjustment. Reimputation methods where the resampling method was jackknife, were suggested by Ford (1983) and Burns (1990) for hot deck imputation. Rao and Shao (1992) show that this overestimates the variance and suggest an adjustment factor. The adjustment suggested by Rao and Shao (1992) is for adjusting

the pseudo value computation in the jackknife procedure. In the pseudo value computation each time a respondent is removed from the sample data, a nonrespondent is adjusted by a factor which is the difference in the means obtained from using all observations and the l^{th} case deleted in the jackknife, i.e.

$$y_i(-l) = y_i + [E(y_i(-l)) - E(y_i)] \tag{6.23}$$

where y_i is a realization in Y, E is the expectation for the random imputation and $y_i(-l)$ is the l^{th} deleted respondent replicate. This adjustment proved to perform well for hot deck, ratio and regression imputation, but does not perform well for the nearest neighbour imputation (Zanutto,1993).

Chen and Shao (2001) observed that computing the variance estimation for nearest neighbour imputation, using jackknife is similar to the situation where jackknife is applied to estimate the variance of non-smooth parameters. For non-smooth parameters even, with complete response jackknife underestimates the variance (Efron, 1994, Shao and Sitter, 1996). Hence they found that direct application of jackknife without allowing for imputation would underestimate the variance. They showed that for nearest neighbour Rao and Shao jackknife overestimates the variance of the parameter of interest. To overcome this difficulty Chen and Shao (2001) gave a correction factor to the Rao and Shao method. The limitation with this method is it needs artificial adjustments. That is in (6.23) the quantity on right hand side must be multiplied by an additional adjustment factor. This adjustment smooths the parameter. Obtaining these adjustment factors may be difficult for complex designs because even for a simple case, deriving these further adjustment factors involve complex mathematical calculations.

Bootstrap methods cope with nonlinear estimates from imputed data such as medians and quantiles. Also we noted previously that Sarndal stated that the method of variance estimation in the presence of imputed values should be practical and

easy to implement and readily accepted by users, which our experience with NFHS-2 data confirms. Therefore we prefer to avoid complex and artificial adjustments, such as that suggested for jackknife by Chen and Shao, for nearest neighbour by using bootstrap methods for estimating V_{tot}.

There are advantages to using bootstrap on the data with nearest neighbour imputation (NNI) (Wang and Shao, 2004).

a) Both NNI and bootstrap are nonparametric,

b) No artificial adjustment as for jackknife is required to estimate the variance correctly

c) It is a unified method for estimating a range of parameters, such as medians, quantiles, means and totals.

d Bootstrap methods are generally easy to implement and avoid the complexities that appeared in jackknife methods.

Unlike in jackknife, in bootstrap we draw samples repeatedly to find the estimate of the parameter of interest. That is, a new set of samples are drawn from the survey sample for finding the parameter of interest. Thus every replicate is a new sample, and is most likely to have a different set of respondents and nonrespondents. Hence, if we do not re-impute the donor may not be in the replicates. Therefore reimputation is the recommended way for estimating the variance when the bootstrap method is used on imputed data.

6.2.3 Summary of the Existing Methods

A summary of various estimating methods correctly available for finding the variance of the parameter of interest are presented in Table 6.1. A tick means the method is

Table 6.1: Summary of the available variance estimating methods

Method	Type of imputation								
	Hot Deck			Model Based			NNI		
	Sample design								
	SRS	Stratified	CD[a]	SRS	Stratified	CD	SRS	Stratified	CD
Bootstrap	√	√[b]	×	√	√	×	×	×	×
Jackknife	√	√	√	√	√	×	√	√	×
Model	√	×	×	√	×	×	√	×	×

[a]Complex designs (Cluster, Multistage)

[b]estimator exists

derived and the estimator is available, and a cross means the method is not derived or verified under that given design and for a defined method of imputation. These methods are for finding the variance of the estimates obtained from the imputed data using single stochastic imputation methods. All methods in Table- 6.1 are where the imputation is performed for one variable and the parameter of interest is a simple mean or total.

In the above table we notice that the variance estimators are available for both resampling and model assisted methods when the survey sample is a simple random sample and the imputation method is simple hot deck method or model based imputation. For complex designs we notice that resampling methods for variance estimation are available for some designs and imputation methods though not all. This is because for sample surveys involving multistage sampling and stratification the calculation of consistent estimates of variance is not a simple task even with complete responses. With missing values there is an added complexity to the estimations. Moreover in imputing the missing values all imputations assume a model

explicitly or implicitly. Deriving the variance estimates under model assisted approaches require an explicit regression model. As a result this can be more sensitive to the deviations from the model assumptions when the imputation model is implicit. Hence in this thesis we do not use model assisted approach. As stated earlier we do not make use of the artificial adjustments as suggested in Chen and Shao (2001). We prefer to use bootstrap methods as they are robust to misspecification of the imputation model (Lee *et al*, 2002). Moreover nearest neighbour is a non-parametric method and so is the bootstrap method. Therefore we will investigate the use of the bootstrap methods for estimating the variance.

6.3 Bootstrap

Bootstrap methods are not readily available for use in surveys like NFHS-2. In addition the bootstrap method has been studied only under hot deck and model assisted imputation, but not for nearest neighbour imputation which is our preferred method of imputation. Hence in this chapter we do some modifications to Shao and Sitter's (1996) bootstrap for the hot deck imputation method and apply it to nearest neighbour method.

We start with the bootstrap for a full response situation and then describe the bootstrap proposed by Shao and Sitter (1996) for finding the variance of the estimator obtained from imputed data. We then present our modification of the Shao and Sitter (1996) method for use in nearest neighbour imputation.

6.3.1 Bootstrap Procedure for Full Response:

Let $Y = (y_1, y_2, ...y_n)$ be a sample of n values drawn from an unspecified probability distribution F. Let θ be the parameter of interest estimated by $\hat{\theta} = g(Y)$. The basic

steps for the bootstrap procedure are (Efron and Tibishirani, 1993)

1. Construct an empirical probability distribution of Y, $\hat{F}$, by giving an equal chance to each unit a sample is being selected. That is from the sample we select each value $y_1, y_2....y_n$ with a probability of $1/n$.

2. Take a simple random sample with replacement of size n from $\hat{F}$,

$$Y^{*b} = (y_1^*, ...y_n^*)$$

3. Corresponding to bootstrap sample Y^{*^b} is a bootstrap replicate of the statistic of interest $\hat{\theta}$

$$\hat{\theta}_b^* = g(Y^{*b})$$

4. Repeat steps 2 and 3 B (where B is large) times. This yields bootstrap samples $Y^{*1}.....Y^{*B}$ and the statistic of interest corresponding to each bootstrap sample

$$\hat{\theta}_b^* = g(Y^{*b}) \; b = 1, 2, ...B$$

5. Estimate the standard error of distribution of $(\hat{\theta})$ using the standard deviation of the B bootstrap samples

$$\widehat{se}_B = \frac{1}{B-1}\left[\sum_{b=1}^{B}(\hat{\theta}^*(b) - \hat{\theta}^*(.))^2\right]^{1/2} \tag{6.24}$$

where $\hat{\theta}^*(.) = \frac{\sum_{b=1}^{B}\hat{\theta}^*(b)}{B}$.

6.3.2 Bootstrap Procedure for Imputed Data

Shao and Sitter (1996) (hereafter SS) give a procedure for estimating the variance of the parameter estimated from data with imputed values created using hot deck

imputation. The procedure is described for the simple case where original survey sample (from which the population estimates are to be computed) is drawn using a simple random sampling. Let Y_{obs} be the r observed cases and Y_{mis} be the m missing cases, $n = r + m$, $Y = (Y_{obs}, Y_{mis})$. Let R be the indicator variable, with $R = 1$ if Y is observed, $R = 0$ otherwise.

1. Draw a simple random sample with replacement of size n from Y. The bootstrap sample (Y^*) will be of size $n = m_b + r_b$, where m_b is the number of nonrespondents and r_b is the number of respondents in the bootstrap sample. Note that m_b is not generally equal to m.

2. Re-impute the missing values in the bootstrap sample Y^*, using the same imputation method as used for imputing Y. This gives an imputed bootstrap sample Y_I^*.

3. Obtain the statistic of interest (θ^*) from Y_I^*

4. Repeat steps 1-3 B times (B large). Use (6.24) to calculate the variance V_{tot} (is $V_{sam} + V_{imp}$).

Although the entire process of doing bootstrap sampling for imputed data is similar to the full response situation, step-2 above makes it different. As noted in step 1, the bootstrap sample may or may not have the same number of the nonrespondents as there were in the original survey sample. This method can have problems if the sample is small and the percentage of nonrespondents is high, though we would question the effectiveness of imputation in such situations. Under this scenario the bootstrap sample could well consist entirely of nonrespondents and thus have no respondents. Then some adjustments are required to the bootstrap sample in order to apply the hot deck imputation method. Another situation where bootstrapping

may be inefficient is where only respondent data is in the bootstrap sample. If this occurs not all bootstrapped samples will have imputed values and the bootstrap samples are not representative of $\hat{F}$. In either of the extreme situations described above, for the data sets with missing values, there is a possibility that the basic assumption of bootstrap (i.e. the bootstrap sample is representative of the survey sample) is not met. For example let us consider the situation of bivariate sample, as in this thesis. Suppose that X is completely observed and Y, has nonresponse, for the above situations there is a possibility that in the bootstrap samples we may have a set of X with no corresponding respondents Y - i.e $m_b = n$, making the bootstrap sample univariate. Another situation is both where both X, Y may be completely observed i.e $m_b = 0$, thus violating the basic assumption of bootstrap. Making $m_b = m$ for all bootstrap samples is desirable as this helps to make all the bootstrap samples representative of the original survey sample (Davison and Hinkley, 1992) Keeping this in mind modifications are made to SS method.

6.4 The Modification of SS Method

Given the problems with SS method we have outlined above, we describe a modification to overcome this. With each bootstrap sample there are three possible scenarios; $m_b = m$, $m_b < m$ and $m_b > m$. For our modification we use Bello's (1994), concept that the bootstrap sample for imputed data should have the same proportion of respondents and nonrespondents as occurred in the original sample. An application of this modified bootstrap method is studied under a simple situation where the survey sample was drawn using a simple random sample and missing values are created using missing at random (MAR) mechanism. For imputation we use nearest neighbour imputation. The performance of the modified bootstrap is

compared with SS method.

6.4.1 Modification of the SS Method in Simple Case

To ensure the proportion of nonrespondents and respondents to be the same proportion as in the survey sample collected we modified SS method as follows. Compute the propensity score as described in section 4.6 for the n cases in the survey sample using the covariates ($\mathbf{X}$) and the response indicator R. Add this score to the survey sample data. From this extended survey sample data we draw a bootstrap sample of size n using simple random sample with replacement. There are three possible outcomes for our bootstrap samples

a) $m_b = m$. That is, the number of nonrespondents in the survey sample and the bootstrap sample are same. Here we simply apply SS method to estimate the variance of the parameter of interest.

b) $m_b < m$ The number of nonrespondents in the bootstrap sample is less than the number of nonrespondents in the survey sample. Hence we need to adjust the bootstrap sample data so that these additional respondents are replaced by nonrespondents. For this we use propensity score.

   1. Sort the bootstrap data using propensity scores.
   2. Divide the data into respondents and nonrespondents
   3. Since a lower propensity score implies a higher probability of being nonrespondent assume the $m - m_b$ respondents with lowest propensity scores to be nonrespondents.
   4. The new bootstrap data set will have three components i.e. the $n - m$ respondents in the original sample, $m - m_b$ respondents forced to be

nonrespondents and m_b nonrespondents.

5. Apply steps 3-5 of SS (p.157) method.

c) $m_b > m$ Here the number of nonrespondents in the bootstrap samples is more than the number of nonrespondents in the survey sample. Hence, we need to adjust the bootstrap sample so that these additional nonrespondents are replaced with respondents. For this we again make use of propensity score.

1. Sort the bootstrap sample data using propensity scores.
2. Divide the data as respondents and nonrespondents
3. since a higher propensity score implies a lower probability of being non-respondent take the $m_b - m$ respondents with highest propensity score and use them to replace the $m_b - m$ nonrespondents with the highest propensity scores.
4. The new bootstrap data set will have three components i.e. the $n - m_b$ respondents in the original sample, $m_b - m$ nonrespondents replaced by respondents and m_b nonrespondents in the bootstrap sample.
5. Apply steps 3-5 of SS method.

Modifying the bootstrap sample in this manner ensures that each bootstrap sample has the same proportion of missingness as in the survey sample for every draw and this allows efficient use of imputation for all bootstrap samples.

6.5 Outline of the Bootstrap Method for Cluster Designs with Imputed Data

In this section we outline a possible extension of bootstrap methods to cluster sample designs with imputed data. Bootstrap sampling outlined in section 6.3.1 can be applied directly to survey sampling when the survey sample design is simple random sample. However when the survey designs are either stratified, cluster or multistage designs the simple bootstrap would underestimate the variance of the parameter of interest even if we had full response. Rao and Wu (1988) derived some modifications which, when applied to bootstrap samples, result in the variance estimator being the same as the theoretical estimator.

For the data with imputed values we make use of the modified bootstrap discussed in section 6.4.1 along with Rao and Wu's (1988) approach for full response cluster designs. For this application we make the following assumptions

- Imputed values come from the same cluster. That is, no respondent from other clusters are used for imputing the missing data within a particular cluster.
- In the sample design, the clusters are selected with equal probabilities and without replacement.
- Cluster sizes are assumed to be the same for both.

For this situation the outline of the bootstrap method is as follows.

1. For the bootstrap sample, we first select, using simple random sample with replacement, K first stage units (e.g village for NFHS-2) from the K first stage units of the survey sample .

2. In the $i^{th}(i = 1, 2...K)$ selected first stage unit from step 1 take a bootstrap sample of size k secondary stage units (SSU) from the k secondary stage units of the bootstrap sample with simple random sample with replacement.

3. once again the samples in the clusters will have the same situation described in modified bootstrap. That is $m_b^k = m^k$ or $m_b^k < m^k$ or $m_b^k > m^k$ where m_b^k is the number of nonrespondents in the k^{th} bootstrap cluster and m^k is the nonrespondents in the k^{th} survey sample cluster.

4. in all cases apply the modified bootstrap outlined in previous section

5. Re-impute missing values in bootstrap sample $Y^\star$, using the same imputation method as used for imputing the missing Y. This gives an imputed bootstrap sample $Y_I^\star$.

6. Adjust each observation in the imputed bootstrap sample as

$$y_{ij}^{adj} = \hat{\bar{Y}}_I + \lambda_i \left[\frac{\hat{Y}_i^\star}{\bar{M}_0} - \hat{\bar{Y}}_I \right] + \lambda_{2i}^\star \left[\frac{M_i^\star y_{ij}^\star}{\bar{M}_0} - \frac{\hat{Y}_i^\star}{\bar{M}_0} \right] \tag{6.25}$$

where $\hat{\bar{Y}}_I$ is the mean obtained from the survey sample imputed data. $\hat{Y}_i^\star$ is the cluster mean obtained from the bootstrap sample $\bar{M}_0 = \frac{1}{K}\sum_{i=1}^{K} M_i$ and M_i is the size of the second stage units in the i^{th} cluster of the population, $y_{ij}^\star$ is the currently selected bootstrap element, $\lambda_1 = \sqrt{(\frac{k}{k-1}(1 - f_1))}$, where $f_1 = \frac{k}{K}$, $\lambda_{2i}^\star = f_1(1 - f_{2i}^\star)$, $f_{2i}^\star = \mathcal{M}_i^\star / M_i^\star$, where $\mathcal{M}_i$ is the size of the SSU selected in the sample. $\mathcal{M}_i^\star$ is the size of the SSU selected in the bootstrap sample and $M_i^\star$ is the corresponding size of the SSU from the population.

7. calculate the parameter of interest θ_b^*

8. Repeat steps 1-7 B times find the parameter of interest $\theta^\star$

9. Find the variance using (6.24).

6.6 Results and Conclusions

For the initial implementation of the modified bootstrap method we start with a simulated data.

Simulation: The population was simulated using a bivariate normal distribution with means $(\mu_1, \mu_2) = (23.24, 2944.27)$. Here μ_1= mean age of the woman and μ_2= mean birth weight of her child. The variance covariance matrix $\Sigma = \begin{bmatrix} 5.30^2 & 348.98 \\ 348.98 & 729.2143^2 \end{bmatrix}$ is obtained from the low birth weight data from Homser and Lemeshow (2002) . A sample of size 1000 was generated.

In the sample thus obtained we created nonresponse in low birth weight using a simple MAR mechanism. To create nonresponse we took a random sample of the covariate. Low birth weight values corresponding to these sampled covariate values were identified as nonrespondents. Once the nonresponse was created we defined a response indicator R, where $R = 1$ if Y observed and $R = 0$ if not. Using the response indicator and the covariate X we find the propensity score as defined in chapter 4 (4.4). To estimate the propensity score we used logistic regression. Once the propensity scores were known we added them to the data.

From the data imputed we obtained the bootstrap samples as outlined in the previous section. For comparison we used SS method. In the comparisons we used different nonresponse rates (5%, 10%, 15%, 25%) and performed 1000 simulations for SS method and our method. These were repeated 400 times. In the bootstrap samples to impute the missing data we used a simple nearest neighbour imputation,

where the nearest neighbour is defined as

$$d_{ij} = |X_i - X_j| \ \forall i \in obs \text{ and } j \in mis$$

The nearest neighbour obtained for the missing case j is the case k for which $d_{kj} = \min_{1 \leq i \leq r}(d_{ij})$.

Results

The results from these comparisons are presented in Table-6.2. In the table we present the coefficient of variation (CV) and the relative bias (RB). The coefficient of variation is computed as the ratio of the standard error (se) and the mean. That is

$$CV = se_{\bar{\bar{y}}_I} / \bar{\bar{y}}_I,$$

where $\bar{\bar{y}}_I = \frac{1}{1000} \sum_{i=1}^{1000} \bar{y}_{Ii}$ and $se_{\bar{\bar{y}}_I} = \sqrt{(6.24)}$ obtained using SS method or our modified method and $\bar{y}_{Ii}$ is the imputed estimator of the mean obtained from the imputed data in the i^{th} bootstrap replicate. The relative bias of $\bar{\bar{y}}_I$ to population mean μ_2 is computed as RB $= bias/\mu_2$ where μ_2 is the population mean and bias is given as

$$bias = \frac{1}{1000} \sum_{i=1}^{1000} (\bar{y}_{Ii} - \mu_2)$$

From Table-6.2 it is observed that the imputed estimator $\bar{y}_{Ii}$ obtained using nearest neighbour imputation is approximately unbiased for the population mean, μ_2.

The relative bias of the variance estimator of $\bar{y}_{Ii}$ obtained from bootstrap (v_B) to the mean squared error of $\bar{\bar{y}}_I$ $MSE_{S\mathcal{R}}$ or V_{tot} (see 6.2) was calculated using

$$\frac{\bar{v}_B - V_{tot}}{V_{tot}}$$

where

$$\bar{v}_B = \frac{1}{400} \sum_{i=1}^{400} v_{B_i} = \frac{1}{400} \sum_{i=1}^{400} var(\bar{y}_{Ii})$$

and

$$V_{tot} = MSE_{SR}(\bar{\bar{y}}_I) = \frac{1}{400}\sum_{i=1}^{400}(\bar{y}_{Ii} - \mu_2)^2$$

A lower relative bias indicates that the bootstrap estimates are close to the total variance $V_{tot}(= V_{sam} + V_{imp})$. From Table-6.3 we see that both Shao and Sitter's method and our modified bootstrap capture the variance due to imputation and performs well as an estimator of variance of $\bar{y}_{Ii}$ for a range of nonresponse rates. It can be inferred that both methods performs well but further analysis is needed to determine if there are cases where there is definitely a difference. However as described earlier, the modified method overcomes the problem with SS that was discussed in section 6.3.2

Table 6.2: Variance, coefficient of variance and relative bias of the data imputed by nearest neighbour imputation

5% nonresponse				
Method	Mean	Variance	CV	RB
SS Method	2925.4	590.9	0.0083	0.001
Modified SS	2971.2	599.9	0.0082	0.001
10% nonresponse				
Method	Mean	Variance	CV	RB
SS method	2926.4	590.5	0.0087	-0.006
Modified SS	2966.9	598.6	0.0085	0.007
15% nonresponse				
Method	Mean	Variance	CV	RB
SS Method	2923.6	591.5	0.0085	-0.007
Modified SS	2969.5	600.9	0.0083	0.008
25% nonresponse				
Method	Mean	Variance	CV	RB
SS Method	2915.6	587.7	0.0082	-0.010
Modified method	2977.0	600.1	0.0083	-0.011

Table 6.3: Relative bias of v_B to empirical MSE

	Nonresponse			
Method	5%	10%	15%	25%
SS Method	-0.07	-0.02	0.01	0.01
Modified SS	-0.06	-0.01	0.03	0.02

Chapter 7

Conclusion

The two main objectives of this thesis were,

1. To develop imputation methods for adjusting for nonresponse in NFHS-2, and

2. Find an approach to estimating the variance of the parameters obtained from the imputed data under the proposed method of imputation.

7.1 Imputation Methods

We have looked at various imputation methods currently available and our overview of these methods is presented in chapter 2. In many survey organizations missing data is imputed using nearest neighbour method. Most nearest neighbour methods described are for a single continuous covariate situation, but the NFHS-2 data contains categorical as well as continuous type of covariates. Usually in such situations, regression based nearest neighbour methods are used. But these make use of a model and the quality of imputation depends on predictive power of the model. To avoid model misspecification, we developed a nonparametric method which makes use of

both categorical as well as continuous variables. The new method is described in chapter 3.

The advantages with this method are that it

- is nonparametric, hence avoids model assumptions.
- deals with asymmetric binary variables.
- handles both categorical and continuous variables.
- makes it possible to accommodate weights for any important covariates based on the subject matter or expert knowledge.
- can be extended to accommodate complex designs.

To evaluate the performance of our method we compare it to the regression based nearest neighbour (RBNN). RBNN is a commonly used method in many survey organizations. We compared the methods with several MAR models. Our comparisons, presented in chapter 5, show that:

- when the response model is a simple MAR, MVNN method generally imputes values closer to the true values than RBNN method.
- when the response models are structured as in simple MAR, results for MVNN are similar to that of RBNN. The mean square error of the imputed values (MSEI) for MVNN method is generally less than that of RBNN.
- for small data sets and high nonresponse rates, one must be careful in using MVNN or RBNN. Here data reduction may be advantageous, though the effectiveness of any imputation technique should be questioned.

- for small sample sizes and low nonresponse levels and for various MAR models, both MVNN and RBNN perform similarly.

- generally MVNN and RBNN perform well but MVNN has several advantages as listed above.

However while MVNN is a good imputation method, it is computer intensive, and so we looked at data reduction methods. Three methods of data reduction were studied. They are graphical modelling, principal component analysis and propensity matching which is a new application of propensity scores as described in chapter 4. The results presented in chapter 5 show that among the data reduction methods studied:

- propensity matching was the fastest.

- no significant differences were noticed among the three data reduction methods.

- generally the MSEI for propensity score was less compared to other data reduction methods.

- there was generally no significant difference between using the reduced data sets and the full data set.

Comparisons between the propensity matching, RBNN and MVNN was also done and reported in Murthy and Chacko (2004) see Appendix-E. In this paper it was observed that

- MVNN was generally better than RBNN and propensity matching, but was computationally intensive.

- Propensity matching was generally similar to matching using MVNN method.
- Propensity matching was generally better than RBNN method especially when the regression model was not a good fit.
- Under certain MAR models, propensity matching was similar to RBNN method.

From these results we recommend MVNN as the generally preferred method for imputation especially when there are variables of mixed types. If computing time is a concern, then data reduction by NNPS may be the preferred choice. The methods proposed in the thesis were tested for different nonresponse rates and MAR mechanisms. Table-7.1 summarizes the applicability of the methods proposed and used in the thesis.

7.2 Variance Estimation

Naive variance estimation methods cannot be used on the data with imputed values as they will under-estimate the total variance. Hence, we looked at currently existing methods for estimating the variance and use the bootstrap method for the following reasons:

- Bootstrap method is nonparametric, and the nearest neighbour method we developed is also nonparametric.
- No complex adjustments as in the jackknife method are required.
- Bootstrap is a unified method for estimating various smooth as well as non-smooth parameters.

We modified the Shao and Sitter's (1996), bootstrap method because their method has problems when applied to the situations explained in 6.3.2, as then the bootstrap

Table 7.1: Summary of the applicability of the imputation methods under MAR response mechanism

MAR type	Data set	Types of Variables	% Nonresponse	NN Method	Reason
Simple	Small	Mixed	5,10,15,25	MVNN	Less bias compared to RBNN and is Nonparametric
	Large	Mixed	5,10,15,25	NNPS	if time constrained else MVNN after data reduction using PCA or Graphical Modelling
Linear	Small	Mixed	5,10,15,25	MVNN	Nonparametric and hence avoids model assumptions
	Small	Continuous	5,10,15,25	RBNN	if regression model is good
	Large	Mixed	5,10,15,25	NNPS	if regression model is poor and computing time is constraint
	Large	Mixed	5,10,15,25	MVNN	after data reduction using PCA or graphical modelling
	Large	Continuous	5,10,15,25	RBNN	if regression model is good. NNPS otherwise
Convex/Concave	Small	Mixed	5,10,15,25	MVNN	If regression model is poor
	Small	Continuous	5,10,15,25	RBNN	If regression model is good. Else NNPS
	Large	Mixed/Continuous	5,10,15,25	RBNN	if model is good else NNPS.

replicates may not be representative of the original sample. Our method keeps the proportion of respondents and nonrespondents in the bootstrap simulations the same as for the original sample . Our modified bootstrap is presented in chapter 6 and indicate that our method generally gives similar estimates of the total variance of the imputed data to Shao and Sitter's method and both give good results with the simulations done so far.

7.3 Future Work

While we have shown that the MVNN works well for the different types of MAR and also for different rates of nonresponse, we have studied that there are a number of areas that needs further study. One area is the derivation of some of the theoretical properties of our method of imputation, similar to the ones developed for the simple nearest neighbour method by Chen and Shao (2000). In chapter 4 of the thesis, graphical modelling has been used as a technique for dimension reduction. This makes use of the causal relation between the dependent variable (Y) and other covariates when reducing the dimension of the data. Attempts to study its efficiency when compared to techniques such as Sliced Inverse regression (Li, 1991) and Principal Hessian Direction (Li, 1992; Cook, 1998) which also make use of information of the dependent variable, will also be studied. Also extension of the variance estimation to complex designs based on the outline described in chapter 6 is required as NFHS-2 is a complex design as well as testing the methods on more data.

Notation

f: sampling fraction

n: sample size (number of observations)

N: population size projected as at December 1998

a: is the number of PSUs to be selected

s: is the number of segments to be selected from a PSU

ℓ: is the subscript for the selected segment

v: number of variables

r: size of nonrespondents

m: size of nonrespondents=$n - r$

$\mathcal{Z}$: data.

$\mathbf{X}$: data matrix of the covariate space $n \times (v)$.

Y: dependent variable $n \times 1$.

$\mathcal{R}$: response indicator, 1 if Y observed.

ψ: an unknown set of parameters corresponding to the response model

d_0^2: Little's MCAR test statistic.

$\mathcal{J}$: number of patterns

r_j= number of respondents in pattern j

obs : observed cases

μ^*= EM estimate of population mean

$\hat{\mathbf{\Sigma}}$: EM estimate of the covariance matrix from observed cases.

d_{ij}= distance between i^{th} and j^{th} unit

$\mathcal{D}$: Mahalanobis distance

mis: missing cases

$\mathbf{X}_{obs}$: covariates corresponding to Y_{obs}.

Y_{mis}: dependent variable with nonresponse

X_{mis}: covariates corresponding to Y_{mis}

θ: an unknown set of parameters corresponding to sampling

$\mathbf{X}'$: (1,$\mathbf{X}$) Covariates with a constant

β: regression parameter

ϵ: regression error term

Y^P: predicted Y

D: dissimilarity

$\mathfrak{m}$: missing case

c: complete case

δ: indicator variable used in dissimilarity computation

ρ_j: range used in dissimilarity computation

I: indicator variable used in dissimilarity computation.

Δ: mean absolute deviation

Γ: Leti's index

$\pi(\mathbf{X})$: propensity score

$\mathcal{E}$: edge of in a graph

$\mathcal{V}$: vertices in a graph

G: graph

$E_{\mathcal{SR}}$: Expectation with respect to (w.r.t) sampling design and response mechanism.

$E_{\mathcal{R}}(.|s)$: denotes the conditional expectation w.r.t response mechanism

$E_{\mathcal{S}}(.)$: denotes expectation w.r.t the sampling

$V_{\mathcal{SR}}(.)$: denotes the joint variance with respect to the sampling and response mechanism.

θ: parameter of interest in general for our study it is mean $\bar{Y}$

$\bar{Y}_I$: Mean obtained from the imputed data

$B_{S\mathcal{R}}$: Bias w.r.t sampling and the response mechanism.

Appendix A

Regions in UP

Table A.1: Regions of the state of Uttar Pradesh

Region	Names of districts
Hill	Nainital, Therigarhwal, Almora, Chamoli, Dehradun, Garhwal, Pittoragarh, Uttarkashi
Western	Bijnor, Ghaziabad, Hardwar, Meerut, Moradabad, Rampur, Sharanpur, Muzzafarnagar, Agra, Aligarh, Bareilly, Budaun, Bulandshahr, Etha, Farrukhabad, Firozabad Mainpuri, Pilibhit, Shahjahanpur, Etawah, Mathura
Central	Kheri, Hardoi, Rae Bareli, Sitapur, Barabanki, Fathepur, Kanpur Dehat, Kanpur Nagar, Lucknow, Unnao
Eastern	Allahabad, Gonda, Pratapgarh, Sultanpur, Bahraich, Faizabad, Azamgarh, Basti, Deoria, Gorakhpur, Jaunpur, Maharajganj, Mau, Siddharthnagar, Ballia, Gazipur, Varnasi, Mirzapur, Sonbhadra
Bundelkhand	Banda, lalitpur, Hamirpur, Jalaun, Jhansi

Source: IIPS, 2000, U.P., state reports

Appendix B

List of Variables

Table B.1: Variables used in simulations

Demographic	**Social**	**Economic and Health**	**Nutritional**
Age	Religion	Current work	Have milk
Children Ever born	Caste	Standard of living	Eat Veges
No. of living children	Education	Suffer asthma	Eat fruits
Ever Terminated pregnancy	Region	Suffer malaria	Eat eggs
Menstruated in last 6 weeks	Chew Tobacco	Suffer Jaundice	Eat Chicken
Current pregnancy	Ever smoke	Suffer jaundice	Green Veges
	Altitude	HL	Eat pulses
	drink alchol		

Appendix C

Non-Response in NFHS-2

Table C.1: Distribution of nonresponse in hemoglobin for NFHS-2 data, by some social, demographic and economic covariates

Covariate	% Nonresponse	%Response
Region		
Hill	61.59	38.41
Western	41.92	58.08
Central	32.96	67.04
Eastern	27.18	72.82
Bundelkhand	55.29	44.71
Residence		
Rural	38.52	61.48
Urban	40.26	59.74

Continued on next page

Covariate	%Nonresponse	%Response
Age		
15-19	41.87	58.13
20-24	39.36	60.64
25-29	35.96	64.04
30-34	40.21	59.79
35-39	37.40	62.60
40-44	38.39	61.61
45-49	40.55	59.45
Education		
Illiterate	40.34	59.66
Literate.< Primary	35.69	64.31
Primary middle	32.98	67.02
Middle school	37.84	62.16
High school	37.59	62.41
Higher secondary	35.02	64.98
Religion		
Hindu	37.75	62.25
Muslim	45.52	54.48
Other	35.80	64.20

Continued on next page

Covariate	%Nonresponse	%Response
Caste		
Scheduled caste	36.31	63.69
Scheduled Tribe	44.50	55.50
Other Backward class	35.77	64.23
Other	41.10	58.90
Currently Pregnant		
No	39.00	61.00
Yes	37.27	62.73
Children ever Born		
0-2	38.18	61.82
2-4	39.34	60.66
4-6	36.61	63.39
6+	38.92	61.08
Current Work		
No	38.53	61.47
Yes	40.03	59.97
Standard of Living		
Low	41.70	58.30
Medium	38.02	61.98
High	35.33	64.67

Bibliography

[1] Agarwal A. K., Sen A. K., Kalra N. K. and Gupta N. Prevalence of anaemia during pregnancy in district burdwan, west bengal. *Indian Journal of Public Health*, 43(1):26–31, 1999.

[2] Allison P. D. *Missing Data.* U7-136. Sage University papers series on quantitative applications in social sciences, Thousand Oaks, CA, 2001.

[3] Bello A. J. A bootstrap method for using imputation techniques for data with missing values. *Biometrical Journal*, 4:453–464, 1994.

[4] Burns R. M. Multiple and replicate item imputation in a complex sample survey. In *Proceedings Sixth Annual Research Conference*, pages 655–665, Washington, D. C.:, 1990. U. S. Bureau of the Census.

[5] Cadima J. and Jolliffe I. T. Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, 22:203–214, 1995.

[6] Cadima J. and Jolliffe I. T. Variable selection and the interpretation of principal subspaces. *Journal of Agricultural Biology and Environmental Statistics*, 6:62–79, 2001.

[7] Cassel C.M., Sarndal C.E. and Wretman J.H. Some uses of statistical models in connection with nonresponse problems. In Madow W.G. and Olkin(eds),

editor, *Incomplete data in Sample Surveys*, volume 3, pages 143–160. Academic Publishers, 1983.

[8] Census of India. *Census year book.* Governement of India, India, 1991.

[9] Cerdeira J. O., Cadima J. and Manuel Minhoto. *The Subselect Package.* [http://cran.r-project.org/doc/packages/subselect.pdf], 2004.

[10] Chen J. and Shao J. Nearest neighbour imputation for survey data. *Journal of Official Statistics*, 16(2):113–131, 2000.

[11] Chen J. and Shao J. Jackknife variance estimation for nearest neighbour imputation. *Journal of American Statistiacal Association*, 96(453):260–269, 2001.

[12] Cochran W.G. and Rubin D. B. Controlling bias in observational studies. *Sankhya A*, 35:417–446, 1973.

[13] Collins L., Schafer J. L. and Kam C. M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330–351, 2001.

[14] D Cook. R. Principal hessian directions revisited(with discussion). *Journal of American Statistical Association*, 93:84–100, 1998.

[15] Cox B. G. and Bonham G. S. Sources and solutions for missing data in the nmcues. In *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pages 444–449, Alexandria, 1983.

[16] Dalenius T. Some reflections on the problem of missing data. In Madow W. G., Olkin I. and Rubin D.B., editor, *Incomplete Data in Sample Survey*, volume 3, chapter 8, pages 411–413. Academic Press, New York, 1983.

[17] Davison A. C. and Hinkely D. V. Computer intensive statistical methods. In Dodge Y. and Whittaker J.(eds), editor, *Compstat*, volume 2, pages 51–62. Physica Verlag, 1992.

[18] Dear R. E. A principal component missing data method for multiple regression models. Technical report, SP-86. Santa Monica, System Development Corporation, 1959.

[19] Deville J. C. and Sarndal C. E. Calibration estimators in survey sampling. *Journal of American Statistical Association*, 87:376–382, 1992.

[20] du Toit S. H. C., Steyn A. G. W. and Stumpf R. H. *Graphical Exploratory Data Analysis.* Springer-Verlag, New York, 1986.

[21] Edwards D. *Introduction to graphical modelling.* Springer, New Jersey, 2000.

[22] Efron B. Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89:463–479, 1994.

[23] Efron B. and Tibshirani R. J. *An Introduction to the Bootstrap.* Chapman and Hall, London, 1993.

[24] de Falguerolles A. and Jmel S. Un critere de choix de variables en analyses en composantes principales fonde sur des models graphiques gaussiens particuliers. *The Canadian Journal of Statistics*, 21:239–256, 1993.

[25] Fay R. E. A design-based perspective on missing data variance. In *Proceedings of the 1991 Annual Research Conference, U.S. Bureau of the Census*, pages 381–440, 1991.

[26] Fay R. E. Comment on multiple imputation inferences with uncongenial sources of input. *Statistical Science*, 9:558–560, 1994.

[27] Felligi I. P. and Holt D. A systematic approach to automatic edit and imputation. *Journal of American Statistical Association*, 71:17–35, 1976.

[28] Ferber Robert. Item nonresponse in consumer survey. *Public Opinion Quarterly*, 30:399–415, 1966.

[29] Ford B. L. An overview of hot deck procedures. In Madow W. G., Olkin I. and Rubin D.B., editor, *Incomplete Data in Sample Survey*, volume 2, chapter 14, pages 206–294. Academic Press, New York, 1983.

[30] Gagnon F., Lee H., Rancourt E. and Sarndal C. E. Estimating the variance of the generalized regression estimator in the presence of imputation for the generalized estimation system. In *Proceedings of the Survey Research Methods Section, Statistical Society Canada*, pages 151–156, 1996.

[31] Godfrey A. J. R., Wood G. R., Ganesalingam S., Nichols M. A. and Qiao C.G. Two-stage clustering in genotype-by-environment analyses with missing data. *Journal of Agricultural Science*, 139:67–77, 2002.

[32] Govidarajulu Z. *Elements of Sampling Theory and Methods.* Prentice Hall, New York, 1999.

[33] Hosmer and Lemeshow. *Applied Logistic Regression.* Wiley, New York, 2002.

[34] Hu Ming-xiu. and Salvucci S. Imputation algorithms. Working Paper 2001-17, National Center for Education Statistics, U.S. Department of Education, 2001.

[35] IIPS. India report. Technical report, International Institute for Population Sciences, Govandi Station Road, Deonar, Mumbai, India, 2000.

[36] Jackson E. J. *A Users Guide to Principal Components.* John Wiley, New York, 1991.

[37] Jolliffe I. T. Discarding varibales in a principle component analysis i: Artificial data. *Applied Statistics*, 21:160–173, 1972.

[38] Jolliffe I. T. Discarding varibales in a principle component analysis ii: Real data. *Applied Statistics*, 22:21–31, 1973.

[39] Jolliffe I. T. *Principal Component Analysis.* Springer, New York, 2nd edition, 2002.

[40] Kalton G. Compensating for missing data. Technical report, ISR research report series, Ann Arbor: Survey Research Center University of Michigan, 1981.

[41] Kalton G. Models in the practice of survey sampling. *International Statistical Review*, 51:175–188, 1983.

[42] Kalton G. and Kaspryzk D. Imputing for missing survey response. In *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pages 22–33, Alexendria, 1982.

[43] Kanani S. Combating anaemia in adolescent girls: A report from india. *Mothers child*, 13(1):1–3, 1994.

[44] Kaufman L. and Rousseuw P. J. *Finding Groups in Data: An introduction to cluster Analysis.* John Wiley, New York, 1990.

[45] Kish L. *Survey Sampling.* John Wiley, New York, 1995.

[46] Korn E. L. and Gourbard B. I. *Missing data in Health surveys.* John Wiley, New York, 1999.

[47] Krzanowski W. J. Cross validatory choice in principal components- some sampling results. *Journal of Statist. Computat. Simual.*, 18:299–314, 1983.

[48] Krzanowski W. J. Selection of variables to preserve multivariate data structure using principal components. *Applied Statistics*, 36:22–33, 1987.

[49] Laaksonen S. Regression based nearest neighbour hot decking. *Computational Statistics*, 15(1):65–71, 2000.

[50] Landerman L. R., Land K. C. and Pieper C. F. An empirical evaluation of the predictive mean matching method for imputing missing values. *Sociological Methods and Research*, 26:3–33, 1997.

[51] Lauritzen S. L. and Wermuth N. Graphical models for association betweeen variables, some of which are qualitative and some quantitative. *Annals of Statisics*, 17:31–57, 1989.

[52] Lee S. E. Imputation using factor analysis. In *Proceedings of the 54th ISI Meetings*, Paper presented at the 54th ISI Meetings, Berlin, 2003.

[53] Lee H., Rancourt E. and Sarndal C. E. Variance estimation from survey data under single imputation. In Groves R. M., Dillman D. A., Eltinge J. L. and Little R. J. A., editor, *Survey Nonresponse*, chapter 21, pages 315–328. John Wiley, New York, 2002.

[54] Lessler J. I. and Kalsbeek W. D. *Nonsampling Error in Surveys.* John Wiley, New York, 1992.

[55] Leti G. Stastisitca descrittiva. I1 Mulino Bologna, 1983.

[56] Li K. C. Sliced inverse regression for dimension reduction (with discussion). *Journal of American Statistical Association*, 86:316–342, 1991.

[57] Li K. C. On principal hessian directions for data visualization and dimension reduction:. *Journal of American Statistical Association*, 87:1025–1040, 1992.

[58] Little R. J. A. Survey nonresponse adjustments for estimate of means. *Internationation Statistical Review*, 54:139–157, 1986.

[59] Little R. J. A. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(4):1198–1202, 1988.

[60] Little R. J. A. Missing data in large scale surveys. *Journal of Business and Economic Statistics*, 6(2):287–301, 1988.

[61] Little R. J. A. and Rubin D. B. *Statistical Analysis With Missing Data.* John Wiley, New York, second edition, 2002.

[62] Little R. J. A. and Schenker N. Missing data. In Arminger G., Clog C. C. and Sobel M. E, editor, *Handbook of Statistical Modeling in the Social and Behavioral Sciences*, pages 39–75. Plenum, New York, 1995.

[63] Little R. J. A. and Smith P J. Multivariate edit and imputation for economic data. In *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pages 518–522, Alexandria, 1983.

[64] Lohr S. *Sampling Design and Analysis.* Duxbury Press, New York, 1999.

[65] Manzari A. and Reale A. Towards a new system for edit and imputation of the 2001 italian population census data: A comparison with the canadian nearest neighbour imputation methodology,. In *Proceedings Actes, of the 53rd International Statistical Institute meetings, Seoul, Invited papers, Volume of the International Association for survey statisticians (IASS.)*, pages 634–655, 2001.

[66] Massawe S. N., Ronquist G., Nystrom L. and Lindmark G. Iron status and iron deficiency anaemia in adolescents in a tanzanian suburban area. *Gynecology Obstetric Invest*, 54(3):137–144, 2002.

[67] McCabe G. P. Principal variables. *Technometrics*, 26:137–144, 1984.

[68] Murthy M. N. and Chacko. E. Imputation by propensity matching. In *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pages [CD–ROM] pp–pp, Alexandria, 2004.

[69] Murthy M. N., Chacko E., Penny Richard and Hossain M. M. Multivariate nearest neighbourhood method of imputation. *Statistics in transition*, 6(1):55–66, 2003.

[70] NCES. *Appendix b: evaluating the impact of imputations for item nonresponse.* National Center For Education Statistics, [http://nces.ed.gov/statprog/2002/appendixb3.asp], 2002.

[71] NFHS-2. National family health survey perliminary reports: Uttar pradesh. Technical report, International Institute for Population Sciences, Govandi Station Road Deonar Mumbai, 2000.

[72] NIS. Anemia. [www.reutershealth.com/wellconnected/doc57.html], 2001.

[73] Oh H. L. and Schuren F. J. Weighting adjustment for unit nonresponse. In Madow W. G., Olkin I. and Rubin D. B., editor, *Incomplete Data in Sample Survey*, volume 2, chapter 13, pages 879–890. Academic Press, New York, 1983.

[74] Platek R. and Gray G. B. Imputation methodology. In Olkin I. Madow W. G. and Rubin D.B., editors, *Incomplete Data in Sample Survey*, volume 2, chapter 17, pages 255–294. Academic Press, New York, 1983.

[75] Ramsay J. O. and Silverman B. W. *Functional Data Analysis.* Springer, New York, 2nd edition, 1997.

[76] Rancourt E. Estimation with nearest neighbour at statistics canada. In *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pages 131–138, Alexandria, 1999.

[77] Rancourt E., Sarndal C. E. and Lee H. Estimation of variance in the presence of nearest neighbour imputation. In *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pages 888–893, Alexandria, 1994.

[78] Rao J. N. K. Jackknife variance estimation with imputed survey data. In *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pages 31–40, Alexandria, 1993.

[79] Rao J. N. K. Varaince estimation with imputed survey data. *Journal of American Statistical Association*, 91:499–506, 1996.

[80] Rao J. N. K and Shao J. Jackknife variance estimation with the survey data under hot deck imputation. *Biometrika*, 79:811–822, 1992.

[81] Rao J. N. K and Sitter R. R. Variance estimation under two phase sampling with application to imputation for missing data. *Biometrika*, 82:453–460, 1995.

[82] Rao J. N. K. and Wu C. F. J. Resampling inference with complex survey data. *Journal of American Statistical Association*, 83:231–241, 1988.

[83] Rizvi M. H. An emprical investigation of soem item nonresponse adjustment procedures. In Olkin I. Madow W. G. and Rubin D.B., editors, *Incomplete Data in Sample Survey*, volume 1, chapter 18, pages 351–381. Academic Press, New York, 1983.

[84] Rosenbaum P. R. *Observational Studies.* Springer, New York, 2nd edition, 2002.

[85] Rosenbaum P. and Rubin D. B. The centeral role of propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[86] Rubin D. B. Inferences and missing data. *Biometrika*, 63:581–592, 1976.

[87] Rubin D. B. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of Amreican Statistical Association*, 72(359):538–543, 1977.

[88] Rubin D. B. Multiple imputation in sample surveys- a phenomenological bayesian approach to nonresponse. In *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pages 20–28, Alexandria, 1978.

[89] Rubin D. B. The bayesian bootstrap. *The Annals of Statistics*, 9:130–134, 1981.

[90] Rubin D. B. Discussion of Hansen, Madow and Tepping. *Journal of the American Statistical Association*, 77:803–805, 1983.

[91] Rubin D. B. The use of propensity scores in applied bayesian inference. In Lindely D.V. Bernardo. J.M., DeGroot M.H. and Smith A.F.M., editors, *Bayesian Statistics*, pages 463–472. Elsevier Science publishers B.V., North Holland, 1985.

[92] Rubin D. B. *Multiple Imputation For Nonresponse in Surveys.* John Wiley, New York, 1987.

[93] Rubin D. B. Imputation in missing data 18+. *Journal Of American Statistical Association*, 91(434):473–489 with discussion 507–515, 1996.

[94] Rubin D. B. and Schenker N. Multiple imputation for interval estimation from simple random sample with ignorable nonresponse. *Journal of Amreican Statistical Association*, 81:366–374, 1986.

[95] Sande I. G. Hot deck imputation procedures. In Olkin I. Madow W. G. and Rubin D. B., editors, *Incomplete Data in Sample Survey*, volume 3, chapter 18, pages 339–349. Academic Press, New York, 1983.

[96] Sapre Shilpa. Anaemia: A most prevalent cause of maternal mortality. *Journal of Obstetrics and Gynaecology of India*, 51(6):23–24, 2001.

[97] Sarndal C. E. Methods for estimating the precision of survey estimates when impuation has been used. In *Proceedings of Statistics Canada Symposium Canda: Measurement and improvement of Data Quality, Statistics Canada*, pages 337–347, 1990.

[98] Schafer J. L. *Analysis of Incomplete Multivariate Data.* Chapman and Hall, London, 1997.

[99] Schieber S. J. A comparison of three alternative techniques for allocating unreported social security income on the survey of the low-income aged and disabled. In *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pages 212–218, Alexandria, 1978.

[100] Shao J. Bootstrap estimation for sample surveys. *Statistical Science*, 17(2):113–131, 2004.

[101] Shao J. and Sitter R. R. Bootstrap for imputed survey data. *Journal of American Statistical Association*, 91(435):1278–1288, 1996.

[102] Shao J. and Steel P. Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of American Statistical Association*, 94(448):254–265, 1999.

[103] Shao J. and Wang H. Sample correlation coefficients based on survey data under regression imputation. *Journal of American Statistical Association*, 97(458):544–552, 2002.

[104] Sixten L. and Sarndal C.E. *Estimation in the presence of nonresponse and frame imperfections.* Statistics Sweden, 2002.

[105] Song J. and Belin T.R. Imputation for incomplete high-dimensional multivariate normal data using a common factor model. *Statistics in Medicine*, 23(19):2827–2843, 2004.

[106] Sood S.K., Ramachandran K., Mathur M., Gupta K., Ramalingaswamy V., Swarnabai C., Ponniah J., Mathan V. I. and Baker S. J. W.H.O. sponsored

collaborative studies on nutritional anaemia in india. 1. the effects of supplemental oral iron administration to pregnant women. *Quantitative Journal of Medicine.*, 44(174):241–258, 1975.

[107] Tanner M. A. and Wong W. W. The calculation of posterior distribution by data augmentation. *Journal of Amreican Statistical Association*, 82:528–550, 1987.

[108] Tollefson M. and Fuller W. A. Variance estimation for sample with random imputation. In *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pages 758–763, Alexandria, 1992.

[109] Tukey J. W. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29:614, 1958.

[110] UNEP report. Children in the new millennium: Environmental impact on health the online version. [www.unep.org/ceh/chapter02.pdf], 2002.

[111] Vacek Pamela. M and Ashikaga Takamaru. An examination of the nearest neighbour rule for imputing missing values. In *Proceedings of the Statistical Computing Section*, pages 326–331, Washington, D.C.:, 1980. American Statistical Association.

[112] Valero P. and Young F. *Missing Data Analysis.* The Visual Statistics System (VISTA), Chapel Hill N.C., 2000.

[113] Wang H. and Shao J. Asymptotic inference based on nearest neighbour imputation and the bootstrap. In *Invited presentation at the joint statistical meetings of the American Statistical Association, Toronto*, 2004.

[114] Westbrooke I. and Jones L. Imputation of maori descent for electoral calculations. Technical Report 01.095.0000, Statistics New Zealand, October 2000.

[115] Whittaker J. *Graphical Models in Applied Multivariate Statistics.* John Wiley, New York, 1990.

[116] Yansaneh I. S., Leslie S. W. and Marker D. A. Imputation methods for large complex datasets: An application to the nehis. In *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pages 314–319, Alexandria, 1998.

[117] Youden W. J. *Statistical Methods for Chemists.* John Wiley, New York, 1951.

[118] Zanutto E. Jackknife variance estimation under imputation for missing survey data. Masters thesis, Carleton University, Ottawa, 1993.

[119] Zio M. D., Scanu M., Coppola L., Luzi O. and Ponti A. Bayesian networks for imputation. *Journal of the Royal Statistical Society*, series A, 167(2):309–322, 2004.

Index

Appendix D

R Functions

```
#**** Computation of Dissimilarity matrix********

## funcion dismt computes the dissimilarities
dismt <- function(data,type,cases=1:nrow(data),wgts=c(rep(1,ncol(data)))){
     ncases <- nrow(data)
     nvar <- ncol(data)
     cms <- length(cases)
     tmpm <- matrix(,ncases,(2*nvar+cms))
     r <- apply(data,2,max,na.rm=TRUE)-apply(data,2,min,na.rm=TRUE)
     for(ii in 1:length(cases)){
         i <- cases[ii]
         for( k in 1:nvar) {
               tmpm[,k] <- 1
```

```
                tmpm[,k] <- !(is.na(data[c(rep(i,ncases)),k]+data[,k]))
                if(type[k]=="a")
                  tmpm[,k] <- !(data[c(rep(i,ncases)),k]==data[,k] & data[,k] ==1)
                if(type[k] == "i" || type[k] == "o" )
                     tmpm[,k+nvar] <- abs(data[c(rep(i,ncases)),k]-data[,k])/r[k]
                if(type[k] != "i" && type[k] != "o" )
                      tmpm[,k+nvar] <- data[c(rep(i,ncases)),k]!=data[,k]
                tmpm[,k+nvar] <- tmpm[,k+nvar]*wgts[k]
               }
tmpm[,(ii+2*nvar)] <- apply(tmpm[,(nvar+1):(2*nvar)],1,sum,na.rm=TRUE)
                           /apply(tmpm[,1:nvar],1,sum,na.rm=TRUE)
          }
tmpm[,(2*nvar+1):(2*nvar+cms)]
 }

#****************************************** End of Dismt function******************

#********* function impute does imputation by using Dismt function*****************

imp <- function(data,type,weights=c(rep(1,ncol(data)))){
    nvars <- ncol(data)
    ncases <- nrow(data)
    miss.vals <- which(is.na(data))
    if(nvars > length(weights) )
                stop("No. of weights is less than No. of variables")
```

```
    if(nvars > length(type) )
              stop("No. of types is less than No. of variables")
    if(length(miss.vals)==0) warning("No Missing Values in Data")
else {
## need to skip out more eleagently
    miss.vars <- (miss.vals-1) %/% ncases + 1
    miss.cases <- miss.vals - (miss.vars-1)*ncases
    miss.dism <- t(as.matrix(dismt(data,type,miss.cases,weights)))
    mins <- if(length(miss.vals)==1)
                  min(miss.dism[,-miss.cases])
             else
                  apply(miss.dism[,-miss.cases],1,min)
    cases <- as.vector(c(1:ncases))
    mv <-  t(miss.dism==mins)
    impv <- vector("logical",min(nrow(mv),ncol(mv)))
    impd <- vector("logical",min(nrow(mv),ncol(mv)))
    if(length(miss.vals)==1) {
          for (i in cases[-miss.cases]) {
                 if(mv[i]) impv <- data[i,miss.vars]
                 if(mv[i]) impd <- i
                }
          }
    if(length(miss.vals)> 1) {
          for(j in 1:ncol(mv))
              for (i in cases[-miss.cases]) {
                 if(mv[i,j]) impv[j] <- data[i,miss.vars[j]]
```

```
                    if(mv[i,j]) impd[j] <- i
                  }
              }
c(length(miss.cases),miss.cases,mins,impd,impv)
 }  ## else to no missing values
 }
#*************************************End of impute ******************


#******************  function miss finds missing values*****************
miss <- function(data){
    nvars <- ncol(data)
    ncases <- nrow(data)
    miss.vals <- which(is.na(data))
    if(length(miss.vals)==0) {
 0
}
else {
    miss.vars <- (miss.vals-1) %/% ncases + 1
    miss.cases <- miss.vals - (miss.vars-1)*ncases
    miss.cases
    }
 }
#*********************************End of Miss function******************

#*********** Calling dismt and imp functions to impute by MV NN method****
```

```
impmv <- function(fl,tpe,mvar=ncol(fl)){
    ra <- imp(fl,tpe)
    nms <- ra[1]
    donda <-fl[ra[(2*nms+2):(3*nms+1)],mvar]
    mv <-ra[2:(nms+1)]
    flmi <- fl
    flmi[mv,mvar] <- donda
  flmi
}
#****************** The gives the fully imputed data***************



#***************************** Distance
function*********************

distance<-function(x,y,z){
    ncase<-length(z)
    don <- 0
   donval <- 0
    for(i in 1:ncase){
        t<-abs(y-z[i])
        m <- x[t==min(t)]
        w <- which(t==min(t))
           if(length(m) > 1.) {
                   w <- sample(w,1,replace=TRUE)
```

```
                    m <- x[w]
                    }
          donval[i] <- m
             don[i] <- w
     }
   return(cbind(donval,don))
}


#***************************** End of distnace function******


#********************Modified regression imputation**************

impr <- function(data){
   nvars <- ncol(data)
   ncases <- nrow(data)
   miss.vals <- which(is.na(data))
   mvars <- (miss.vals-1) %/% ncases + 1
   miss.cases <- miss.vals - (mvars-1)*ncases
   nomiss <- length(miss.cases)

#********************** End of known values******************
     dd<-fac(data)
       mm<-model.matrix(~Age+RU+Region+Religion+
         Caste+CEB+Currpreg+Menlsw+Nlchil+Eveterpreg+Curwrk+Milcurd+
```

```
      Pulbean+Greenveg+Otherveg+Fruit+Eggs+Chickmeat+SSLI+Suffasth+
     Sufftb+Suffmal+Suffjau+Chetob+Drnkal+Smoke++logalt+SEDUC6,dd)
     splm<-lm(Hemoglobin~.,data=dd,na.action=na.omit,singular.ok=TRUE)
    cof<-as.matrix(coef(splm))
      e<-resid(splm)
     el<-length(e)
    df<-df.residual(splm)
    rsd<-sd(e)*sqrt((el-1)/df)
    epi<-rnorm(ncases,0,rsd)
    null.coe<-miss(cof)
     if(!(null.coe==0)){
      cof<-cof[-null.coe]
     mm<-mm[,-null.coe]
}
     prd<-as.vector(t(cof)%*%t(mm))
    prde<-prd+epi
   psim<-cbind(data,prde)
   yobs<-psim[,mvar]
  yhat<-psim[,nvars+1]
   z<-psim[miss.cases,nvars+1]
  imp.val<-distance(yobs,yhat,z,miss.cases)

  c(nomiss,miss.cases,z,imp.val[(nomiss+1):(2*nomiss)],imp.val[1:nomiss])
}
```

```
#************* End of Mdf reg***************

#****** imputation by regression method***********
imprg<-function(flma,mvar){
                ra <- impr(flma)
                nms <- ra[1]
                donda <-ra[(3*nms+2):(4*nms+1)] #donated missing values
                mv <-ra[2:(nms+1)]
                flmi <- flma
                flmi[mv,mvar] <- dond
                flmi
   }
#********************************** END of Reg IMP*****************************



#*******************************Imp by propensity scores*****************
impp<-function(data,mvar){
          dprop<-data[,-mvar]
          nr<-nrow(data)
          r<-rep(1,nr)
          mv<-which(is.na(data[,mvar]))
          r[mv]<-0
          dprop<-cbind(dprop,r)
          rglm<-glm(r~.,data=dprop,family=binomial(logit))
          prd<-as.vector(fitted(rglm))
```

```
            datp<-cbind(data,prd)
            yr<-prd[-mv]
            zr<-prd[mv]
            pt<-proc.time()
            x<-data[,mvar]
            x<-x[-mv]
            disr<-distance(x,yr,zr)
            donvalr<-disr[,1]
        data[mv,mvar]<-donvalr
        ptp<-proc.time()-pt
        impd<-data
        return(impd)
} #*******ENd of propensity matching********

#*****Function to define nominal variables in data**********
fac<-function(data){
     fv<-c(3,4,5,12,13,14,15,16,17,18,29)
     l<-length(fv)
  for(i in 1:l){
    data[,fv[i]]<-as.factor(data[,fv[i]])
  }
    data
} #***********Cumulative function Leti's index****
cf<-function(data){
      ncases<-(nrow(data)-1)
         dat<-0
```

```
  for(i in 1:ncases){
       dat[1]<-data[1]
      dat[i+1]<-dat[i]+data[i+1]
  }
   return(dat)
  }
#**************Function for inducing missingness************

misd<-function(data,qvar,mvar){
        rnm<-as.numeric(rownames(data))
        temp<-cbind(rnm,data)
        for(j in 1:4){
        n<-length(which(temp[,qvar]==j))
        k<-c(0.2,0.4,0.4,0.2)
        r<-round((n*k[j]),digits=0)
            smp<-temp[sample(which(temp[,qvar]==j),r,replace=TRUE),]
            mis<-as.vector(smp[,1])
             temp[mis,mvar]<-NA
  }
            return(temp)
  }


#******************End of function for creating missing data*******************
#******************Program used in the gm analysis.************************
```

```
pcor1<-function(x){
        vc<-var(x)
        inv<-solve(vc)
        d<-1/sqrt(diag(inv))
        d<-matrix(d)
        td<-dd%*%t(dd)
        corr<--(inv*td)
        diag(corr)<--diag(corr)
        return(corr)
}



#****test for selecting edges in gm analysis.********
partest2<-function(data,N){ #data partial correlations

        #lt<-lower.tri(data)
        #vals<-data[lt==1]
        #vals<-as.data.frame(vals)
        chites<--N*log(1-data^2)    #N sample size
        pval<-round(pchisq(chites,1,lower=FALSE),4)
        return(chites,pval)
    }

#**End of defining the nominal variables in data**
```

```
#*****Variance estimation Saho Sitter' method********
#*************************Saho's program*******************

sspr<-function(data){
        nr<-nrow(data)
        I<-which(colnames(data)=="I")
        dats<-data[c(sample(1:nr,nr,replace=TRUE)),]
         dats[dats[,I]==0,2]<-NA
             m<-which(is.na(dats[,2]))
           x<-dats[-m,2]
           y<-dats[-m,1]
            z<-dats[m,1]
          sst<-distance(x,y,z)
            ssimp<-dats
             donv<-sst[,1]
          ssimp[m,2]<-donv
           return(ssimp)
}
#***** End of Saho Sitter method***
#*************************Our Method**********************

pssn<-function(dat){

         s<-nrow(dat) #inital sample size of the data
        m<-which(colnames(dat)=="y")
         mv<-which(is.na(dat[,m]))
```

```
        mvl<-length(mv)
        datr<-dat[-mv,]

       dats<-dat[sample(1:s,s,replace=TRUE),]
       sdats<-dats[sort.list(dats$pr),]
       mvs<-which(is.na(sdats[,m]))
        mvls<-length(mvs)
        datsr<-sdats[-mvs,]
        datsnr<-sdats[mvs,]
         ns <- nrow(datsr)
       nd <- sdats
if(mvls<mvl) {
         datsr[1:(mvl-mvls),m]<-NA
         nd <- rbind(datsr,datsnr)
   }
if (mvls > mvl) {
   datmns<-datsr[(ns-(mvls-mvl)+1):ns,]
   nd <- rbind(datsr,datmns,datsnr[(1:mvl),])
   }
nd
}

#####################End of creating data with same proportions of r and m**
 pss<-function(dat){
 sdats<-pssn(dat)
  mrs<-which(is.na(sdats[,2]))
```

```
                mrsl<-length(mrs)
 #imputing the missing cases with NN method
      x<-sdats[-mrs,2]
        y<-sdats[-mrs,1]
         z<-sdats[mrs,1]
      sst<-distance(x,y,z)
   ssimp<-sdats
    donv<-sst[,1]
     ssimp[mrs,2]<-donv
    return(ssimp)
 }
#**********End of Modified bootstrap*********
```

Appendix E

Published Papers

STATISTICS IN TRANSITION, June 2003
Vol. 6, No. 1, pp. 55–66

MULTIVARIATE NEAREST NEIGHBOURHOOD METHOD OF IMPUTATION

M.N.Murthy[1], Easaw Chacko[1], Richard Penny[2], M.M. Hossain[3]

ABSTRACT

Hot deck imputation procedures are often used to replace missing values. Among hot deck procedures, nearest neighbour methods are preferred over other single imputation procedures because of their efficiency, and asymptotic unbiasedness. We describe a procedure for finding the nearest neighbour when there are many variables that can be of different types, such as binary, nominal, ordinal, ratio and interval. The results obtained by our method are compared with the regression based nearest neighbour method and show that our method generally performs better than regression based nearest neighbour.

Key words: Nearest Neighbour, Regression based nearest neighbour, multivariate nearest neighbour, Imputation and dissimilarity

1. Introduction

In the last thirty years, various procedures have been developed in order to compensate for item nonresponse in surveys. Case deletion and imputation are the common procedures used to deal with item nonresponse. In case deletion, the cases with nonresponse are deleted, and only those cases with a complete set of responses are used for analysis. Case deletion often leads to loss of information and can create biases because systematic differences may exist between respondents and non-respondents.

Imputation is where values are assigned to missing items. There are many methods of imputation (Little and Rubin, 1987; Lohr, 1999; Govindarajulu, 1999; Rao, 2000). A common imputation procedure is hot deck, or real donor, imputation where a complete case (donor) is chosen from the current survey to

[1] University of Canterbury, Christchurch, P.O Bag 4800, New Zealand.
[2] Fidelio Consultancy 264 Grahams Road, Christchurch, New Zealand.
[3] Christchurch School of Medicine and Health Sciences, Christchurch, New Zealand.

supply the values for the missing information in the incomplete case (recipient) (Sande, 1983). Hot deck imputation is a very generic term that has been used for various methods of identifying the "best" donor for the missing values; one of these is the nearest neighbour (NN) method. In the NN method a metric is defined to measure the "distance" between a recipient and all potential donors. The distance measure is computed using variables (covariates), which are not missing for the recipient and potential donors. The respondent "nearest" to the nonrespondent is used as the donor for imputation purposes (Little and Rubin, 1987, p.61). Chen and Shao (2000) discuss the theoretical properties of the NN method when the distance measure is Euclidean and there is only one covariate. Possible ways to find a distance measure for NN methods where the covariates are multivariate have been mentioned in Chen and Shao (2001) but no specific methods are suggested. In this paper, we develop a method for selecting the nearest neighbour using multivariate covariates with possibly mixed types of variable such as nominal, ordinal, binary, and interval.

This paper assumes that the missing data is missing at random (MAR), that is, missingness may depend on the observed variables but not on the variable that we are imputing (Rubin, 1976). Moreover we assume that nonresponse is in one variable. However our work can be extended to multiple item nonresponse. In such situations imputation can be done using the principles described in Little and Rubin (1987, p.17).

The rest of this paper is divided into four sections. In section 2, we describe the National Family Health Survey-2 (NFHS-2) data set, and how this data was used in our simulations. The imputation methods we compare and the measures we use to compare the methods are described in Section 3. In section 4, we present our results and in the final section provide our conclusions.

2. NFHS-2 Data

NFHS-2 is a Demographic and Health survey conducted in India during 1998-1999. The survey has three parts:

1. Household Interview
2. Women's Interview
3. Health Investigation

The first part is the interview of a household in the selected sample. The second part identifies women aged between 15-49 in the household who have been ever married and collects their socio-demographic and economic data. In the final part, the women interviewed in the second part and their children aged less than 3 years of age have their height, weight, and haemoglobin levels (HL) measured. In fact measuring HL in the above age group was one of the major objectives of the survey as it is a prime reason for high maternal mortality in

India. For details of the sample design and survey instruments see NFHS-2 (2000).

2.1. Simulation

The data for our simulation is taken from NFHS-2 data for the state of Uttar Pradesh (U.P). The sample size in U.P. is 9292. In this paper we examine the item nonresponse for HL only, because of the above mentioned reason. Some of the reasons postulated for nonresponse in HL (e.g. women's health) suggest that the missingness is not MCAR.

We use the complete cases as our simulation population (***P***) to study the imputation methods in this paper. Multiple samples of sizes 100, 500, and 1000 are selected from data set ***P***. We then generate item nonresponse for HL assuming MAR; HL is considered missing if $[C_s=1 \cap C_p=1 \cap U< 0.2]$, where C_s and C_p are the covariates caste and current pregnancy respectively, and U is an independent random variable uniformly distributed over [0,1]. Figure1 shows the details of this process.

3. Methods and Materials:

First we introduce the terminology for the purpose of describing the imputation methods, let

Z: sample data matrix (n×p) drawn from ***P***.

X: co-variate matrix (n×p-1)

Y: response variable vector (n×1)

Y has missing observations, therefore $Y = (Y_{obs}, Y_{mis})$. Similarly, the set of covariates corresponding to Y_{obs} (Y_{mis}) is denoted by X_{obs} (X_{mis}).

3.1. Regression Based Nearest Neighbour method (RBNN)

In choosing the best donor for imputation there have been several approaches for defining a suitable distance function for multivariate covariates. One way is to make use of the nearest neighbour hot decking method and a multivariate regression method (Laaksonen, 2000). Such methods are referred to as semi parametric methods (Allison, 2001). Based on the type of response variable *Y*, different regression models are selected. For example, when *Y* is binary then a logistic regression model is used. As our response variable is an interval variable we use linear regression. In RBNN imputation (Laaksonen, 2000) the steps are

a. Fit a model using the complete respondents

$$E(Y_{obs}) = X_{obs}\,\beta$$

b. Use this estimated β ($\hat{\beta}$) to predict vector *Y*

$$Y^P = (Y^P_{obs}, Y^P_{mis}) = \mathbf{X}\ \hat{\boldsymbol{\beta}} + \varepsilon$$

c. Find the nearest neighbour to Y^P_{mis} from Y^P_{obs} for each missing case

d. Use the Y_{obs} corresponding to the nearest Y^P_{obs} as the imputed value for Y_{mis}

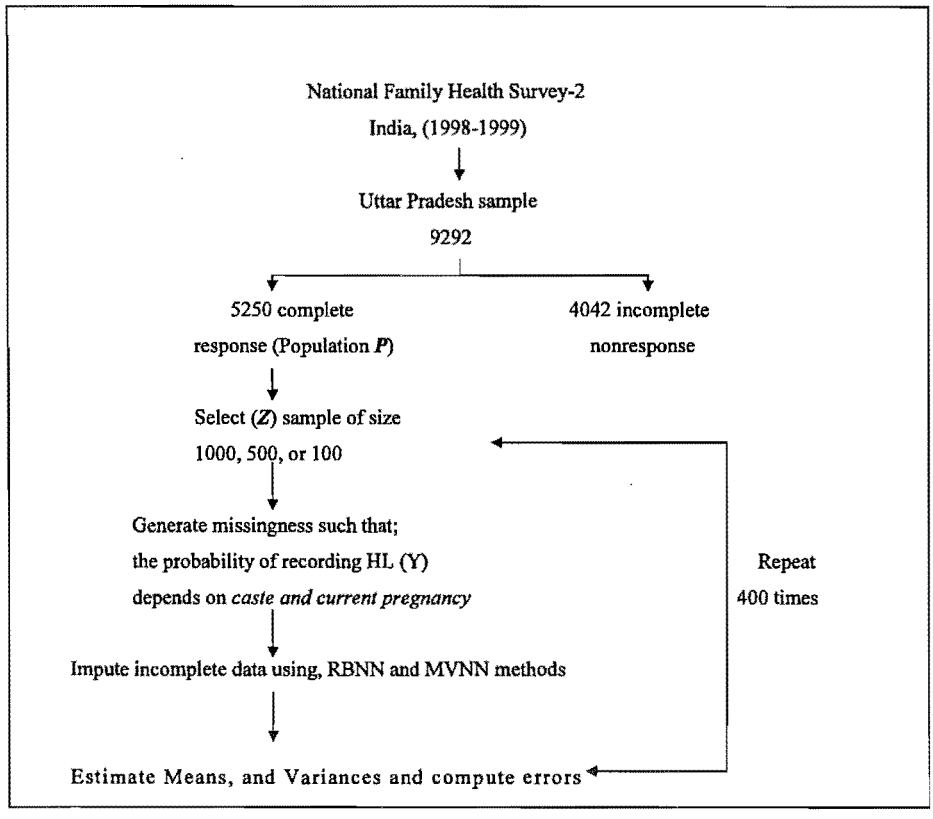

Figure 1. Schematic representation of the simulation procedure

e. Fit a model using the complete respondents

$$E(Y_{obs}) = X_{obs}\,\boldsymbol{\beta}$$

f. Use this estimated $\boldsymbol{\beta}$ ($\hat{\boldsymbol{\beta}}$) to predict vector Y

$$Y^P = (Y^P_{obs}, Y^P_{mis}) = \mathbf{X}\ \hat{\boldsymbol{\beta}} + \varepsilon$$

g. Find the nearest neighbour to Y_{mis}^{p} from Y_{obs}^{p} for each missing case

h. Use the Y_{obs} corresponding to the nearest Y_{obs}^{p} as the imputed value for Y_{mis}

The error (ε) is assumed normally distributed (0, σ^2). It is not uncommon in applications to assume $\sigma^2 = 0$, but this does not add variability due to imputation model. Thus the statistical significance of any analysis will be over-estimated. If the data sets are small and the missingness is large then choice of error term can make a difference to the final data (Allison, 2001). There are several ways to include the error term to the model (Kalton and Kaspryzk, 1981), but we follow Laaksonen (2000) where σ^2 is the mean square error estimated from the regression model.

According to Laaksonen the advantages of RBNN imputation are; it is likely to perform better than simple regression and hot deck imputation; it does not tend to underestimate variance as it is stochastic; and it ranks the respondents and non-respondents so that the probability for nonzero imputed value will be increased while the predicted value will be increased. Similar to all model based imputation methods, RBNN imputation can be problematic for the following reasons;

- Selecting a good regression model can be difficult particularly if there are various types of covariates.
- Outliers will affect parameters estimates and thus imputed values.
- For small sample size and large number of variables there may be a problem of singularity when computing β.
- Failure of multivariate normality assumption may lead to heavy tailed distribution of the imputed values (Schafer, 1997).

3.2. Multivariate Nearest Neighbour Method

Given the problems with RBNN, we have developed an alternate NN imputation procedure using a "distance" function (called dissimilarity) given by Kaufman and Rousseuw (1990) where this is used for cluster analysis. This dissimilarity easily handles mixed types of covariates.

The dissimilarity between a complete case c and a missing case m is defined as

$$D(c, m) = \frac{\sum_{j=1}^{p-1} \delta_{cm}^{j} d_{cm}^{j}}{\sum_{j=1}^{p-1} \delta_{cm}^{j}}$$

where the distance is

$$d_{cm}^{j} = d_{cm}^{j}(x_{cj}, x_{mj}) = \begin{cases} \left.\begin{matrix} 1, & \text{if } x_{cj} \neq x_{mj} \\ 0, & \text{otherwise} \end{matrix}\right\} & \text{when the } j^{th} \text{ covariate is binary or nominal} \\ \dfrac{|x_{cj} - x_{mj}|}{R_j} & \text{when the } j^{th} \text{ covariate is interval.} \end{cases}$$

δ_{cm}^{j} is an indicator variable which takes 1 for all variables except where the jth covariate is asymmetric and $x_{cj} = x_{mj} = 0$. R_j is the range of the jth covariate. We use range rather than standard deviation to normalise so that the interval and ordinal variables have a distance in [0,1], consistent with other variables. Ordinal and ratio scaled variables are treated as interval scaled variables.

An asymmetric binary variable is one where the outcomes are not of equal importance; the more important outcome is coded as 1 and the other as 0, i.e. 1 has more importance than 0. For example, in a symmetric binary variable, say where sex = '1' is male and '0' is female, the categories 1,0 are of equal importance for dissimilarity calculation. However, if we consider the variable "ever terminated pregnancy", where 1 corresponds to "never terminated" and 0 corresponds to "ever terminated". Here "ever terminated" could include terminations occurred recently. These clearly are more important than a termination in the remote past. Hence the 1-1 asymmetric variable pair is a stronger match and more significant than the 0-0 pair in terms of choosing a donor. As the 0-0 pair does not provide useful information for matching it is disregarded in the computation (Kaufman and Rousseeuw, 1990, p. 26).

Of possible donors the case with *min* [$D(c, m)$] is considered the nearest neighbour and used as the donor. If there are several donors with the same *min* [$D(c, m)$], a donor may be randomly selected from among them. Note that $D(c, m)$ has all the properties of a metric except for the triangle inequality. Thus it is not a distance metric as is usual for nearest neighbour methods.

3.3. Measures of Efficiency

In simulation studies one has the advantage of knowing the true values of missing data. Hence in such simulation studies the efficiency of an imputation can be assessed by

1. Comparing imputed values to true values
2. Comparing marginal distribution for data completed by imputation with that of the true data
3. MSE due to imputation for the parameter estimates

3.3.1. Evaluation of Imputation

Evaluation of individual imputations for numeric variables is measured by mean absolute deviation (Manzari and Reale, 2002)

$$\Delta = \frac{\sum_{i=1}^{n_{imp}} | y_i^{imp} - y_i^{act} |}{n_{imp}}$$

where y_i^{act} is the value from **Z**, y_i^{imp} is the imputed value and n_{imp} is the number of imputed values.

3.3.2. Evaluation of imputed marginal distributions

For evaluating the imputed marginal distributions we use an index. This computes the difference between the relative distributions of imputed values and the actual values. The index is (Leti, 1983)

$$\Lambda = \left(\sum_{i=1}^{k} | g(i) - h(i) | / 2 \right) * 100$$

where $g(i)$ is the cumulative relative frequency of the i^{th} category in the imputed data, $h(i)$ is the cumulative relative frequency in the actual data and k is the number of categories. The numeric variables are categorized before computing Λ and the limits for Λ are (0-100). The smaller the value of the index the more the distributions are similar (Manzari and Reale, 2002).

3.3.3. Mean Square Error in Parameter estimates

In order to quantify differences in parameter estimates (e .g. means, variance) for the actual and imputed data we use mean squared errors given by

$$\frac{1}{M} \sum_{s=1}^{M} (\hat{\theta}^{*s} - \hat{\theta}^{s})^2$$

where for each simulation s, $\hat{\theta}^{s}$ is the parameter estimate obtained from **Z**, and $\hat{\theta}^{*s}$ is the parameter estimate obtained from the imputed data.

4. Results

RBNN and MVNN methods have been applied to the data described in section 2. The sampled data (**Z**) is obtained from P and the parameter estimates were computed. The missingness was then generated in the data as described in section 2. For RBNN method the dependent variable Y is HL and the covariate X is a matrix of 28 covariates. The variables in X were selected on the basis of the

literature on the factors that affect HL (UNEP Report, 2002; Shilpa Sapre, 2001; Massawe, 2002; NFHS-2, 2000). Details of some of the variables are listed in Appendix A (full information can be provided on request). The results for the RBNN model show that R^2 was between 0.25-0.72 for the 400 simulations of sample size 100. For MVNN imputation we use the same data and compute the dissimilarities between the covariates of missing cases and cases with no missingness.

Table 1 presents the results obtained for the measures of efficiency described in section 3.3.1, and 3.3.2. The data used for this assessment had sample sizes 1000, 500 and 100 and each of the 400 samples had 20% missing in HL. The results indicate that the mean absolute deviation between imputed and original values is similar for both methods with that for MVNN being consistently smaller than for RBNN.

Table 1. Evaluating individual imputations

Imputation method	Mean absolute deviation between imputed and actual values		
	Samples of sizes		
	1000	500	100
MVNN	20.64	21.3	26.2
RBNN	22.90	26.5	28.0

In order to apply Leti's test we had to categorize our continuous variable HL. We adopted the similar categorization ("Severe", "Moderate", "Mild", "Normal") as in NFHS-2 (2000). From the results in Table 2, we observe that MVNN preserves the marginal densities better than RBNN.

Table 2. Comparing the marginal distributions

Imputation method	Leti's test		
	Samples of sizes		
	1000	500	100
MVNN	0.9	1.5	3
RBNN	2.1	1.9	4.5

Table 3 shows that MVNN parameter estimates has smaller mean squared error compared to RBNN. Figure 2 compares the box plots in 2 parameter estimates obtained under MVNN and RBNN methods for all the simulations for sample size 1000. Figure 2a, comparing the errors in means, show that MVNN is slightly more biased than for RBNN but has a much smaller spread. Figure 2b compares errors in variance and shows MVNN is less biased, less right skewed, and less long tailed than RBNN. Overall this shows that MVNN imputed data has lower MSE than for RBNN. Box plots for samples of 500 and 100 show similar results.

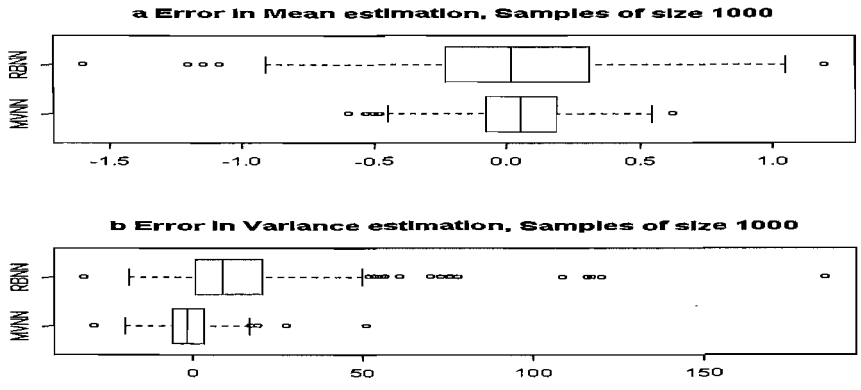

Figure 2. Error in estimates of the parameters mean, and variance

Table 3. Comparison of mean square error for MVNN and RBNN.

Sample Size	Imputation Method	MSE mean	MSE variance (x 10^3)
1000	MVNN	0.14	0.16
	RBNN	0.38	1.94
500	MVNN	0.27	0.29
	RBNN	0.56	2.85
100	MVNN	1.18	1.82
	RBNN	2.72	12.47

Table 4. Mean Square error for varying nonresponse levels for sample size of 100

Parameter estimate	Imputation Method	Nonresponse levels		
		15%	20%	30%
Error in mean	MVNN	0.43	1.18	2.15
	RBNN	0.81	2.72	8.49
Error in variance $\times 10^3$	MVNN	0.42	1.82	3.14
	RBNN	11.70	12.47	65.90

We provide in Table 4 a comparison of the two imputation procedures for varying nonresponse levels (15%, 20%, and 30%), which resemble low, moderate and high nonresponse levels for official statistics. The 30% nonresponse rate is similar to the item nonresponse rate for HL in Uttar Pradesh. Table 4 shows that the MSE increases for both the methods with an increase in nonresponse as would be expected, but the MSE's for MVNN are consistently smaller than for

RBNN. These results support our belief that MVNN is a better method of imputation than RBNN.

5. Conclusions

Hot deck imputation is a common tool for filling in missing values for item nonresponse with NN methods commonly used to find donors. We have presented an improved method of finding the nearest neighbour (MVNN). Our results indicate that the MVNN method performs better than RBNN for finding suitable donors. MVNN imputes the individual records such that marginal distributions are better preserved and the parameter estimates are closer to the true estimates. Our procedure is nonparametric and thus does not require construction of a suitable model.

However MVNN can be computationally intensive with large data sets either of size or number of variables. Therefore we are investigating

1. Condensing the data in order to reduce the computational burden, yet preserving the advantages of MVNN.
2. Weighting covariates in order to reflect subject matter input as to their importance of the variable.
3. Sensitivity analysis for the MVNN method to see if there are breakdown points (eg. Levels of nonresponse, mixture of covariates).
4. Results for different MAR procedures.

Acknowledgements

This research is supported by New Zealand Agency for International Development (NZAID) Scholarship and University of Canterbury Doctoral Scholarship. The authors would like to thank the referees and the editor for their valuable comments and suggestions.

Appendix A: Partly list of variables used in the RBNN and MVNN methods

Variable description					
Variables	**Type**	**Descriptive statistics**			
		Min	**Max**	**Mean**	**Median**
Age	Ordinal	1	7		4.000
RU (Rural & Urban)	Binary	1	2	1.8	2.000
Region	Nominal	1	5		3.000
CEB (Children Ever Born)	Interval	0	13	3.565	3.000
Currpreg^{+} (Current Pregnant)	Binary	0	1	0.08	
Nlchil (Number of living Children)	Interval	0	12	2.957	3.000
Eveterpreg (Ever terminated pregnancy)	Asymmetric Binary	0	1	0.22	
HL g/dl * (Haemoglobin gram/deci litre)	Interval	1.8	31.1	11.78	12.00
Logalt (Altitude log transformed value)	Interval	3.176	4.000	3.965	4.000

* Response variable
\+ Covariates used for conditioning nonresponse.
For further details on the list of all the variables in the survey see to http://www.nfhsindia.org// the data sets can be obtained from http://www.measuredhs.com//

REFERENCES

ALLISON, P.D. (2001) *Missing Data,* Sage university papers series, Quantitative applications in the Social Sciences, Series 136, Thousand Oaks, CA.

CHEN, J. and SHAO, J. (2000) Nearest neighbourhood imputation for survey data. *Journal of Official statistics*, vol.16, 113-131.

CHEN, J. and SHAO, J. (2001) Jackknife variance estimation for nearest neighbour imputation. *Journal of American Statistical Association*, Vol. 96, no. 453, 260-69.

GOVINDARAJULU, Z. (1999), *Elements of Sampling Theory and Methods*, New Jersey: Prentice Hall

KALTON, G. and KASPRYZK, D. (1981) Imputing for survey responses. *Proceedings of the section on survey research methods, American Statistical Association.* 22-33.

KAUFMAN, L. and ROUSSEEUW, P. J. (1990) Finding Groups in Data. *An introduction to cluster analysis.* John Wiley. New York.

LAAKSONEN, S. (2000) Regression-based nearest neighbour hot decking, *Computational statistics*, vol.15, no.1, 65-71.

LETI, G. (1983) *Statisitca descrittiva*, Il Mulino, Bologna.

LITTLE, R.J.A. and RUBIN, D.B. (1987) *Statistical analysis with missing data.* New York: Wiley.

LOHR, S. (1999), *Survey sampling*, New York: Addison and Huxley publications.

MASSAWE, S.N., RONQUIST, G., NYSTROM, L., and LINDMARK, G. (2002) Iron status and iron deficiency anaemia in adolescents in a Tanzanian suburban area, *Gynecology Obstetric Invest* Vol. 54, no.3, 137-144

MANZARI, A. and REALE, A. (2001) Towards a new system for edit and imputation of the 2001 Italian population census data: a comparison with the Canadian nearest neighbour imputation methodology, *Proceedings Actes, of the 53rd International Statistical Institute meetings, Seoul, Invited papers, Volume of the International Association for survey statisticians (IASS.)*, PP-634-655.

National Family Health Survey-2 (2000): International Institute for population Sciences, *National Family Health Survey-2* state reports, Uttar Pradesh, [www.nfhsindia.org].

RAO, P.S.R.S (2000) *Sampling methodologies with applications,* London: Chapman and Hall/CRC.

RUBIN, D.B. (1976) Inference and missing data, *Biometrika* vol. 63,581-90.

SANDE, I. G. (1983) Hot –Deck imputation procedures, in *Incomplete data in sample surveys.* Vol.3. Proceedings of the symposium. W.G.Medow, and I.Olkin (eds). New-York: Academic press. 334-350.

SCHAFER, J. (1997) *Analysis of incomplete multivariate data*, London: Chapman and Hall, 380-386.

SHILPA SAPRE. (2001) Anaemia: A most prevalent cause of maternal mortality, *Journal of Obstetrics and gynaecology of India*, Vol.51, no. 6, 23-24.

UNEP Report. (2002) Children in the New Millennium: Environmental Impact on Health *The Online Version*, Chapter 2 [www.unep.org/ceh/chapter02.pdf].

Imputation by Propensity Matching

Murthy. N. Mittinty* Chacko. E†

Abstract

In large-scale surveys item nonresponse is a common phenomena. Many survey organizations use imputation to deal with missing data. Nearest neighbour imputation (NNI) has gained a lot more attention than other single imputation methods. However, in multivariate covariate situations, finding the nearest neighbour can be complicated when many variables need to be matched. In this paper we show a new application of the propensity score, which we call the nearest neighbour by propensity score (NNPS), for finding a donor for a recipient in multivariate situations. Propensity matching was originally used by Rosenbaum and Rubin (1983) in observational studies. We use propensity score for matching as it assures that the conditional distribution of the covariates given the propensity score is the same for the donors and recipients. NNPS is investigated using simulations assuming that the missing data is either missing at random (MAR) linear or MAR convex. We compare NNPS, with regression based nearest neighbour (RBNN) imputation and a new imputation method given by Murthy *et-al* (2003) called the nearest neighbour by dissimilarity (NNDM). The results indicate that matching by propensity scores seems to be a good choice for many situations, and has the advantage that it reduces the "curse of dimensionality".

Keywords. Propensity Score, Nearest Neighbour, Imputation, Covariate.

1 Introduction

In surveys, item nonresponse is a very common phenomena. Imputation is the common tool used to compensate for the missing data (Rubin, 1987; Chen and Shao, 2000, Rancourt, Sarndal, and Lee, 1994). Single or multiple data sets are created using an imputation technique. Many statistical organizations prefer single imputation to multiple imputation, in order to avoid problems caused by the multiple data sets (Marker *et-al*). In single imputation, there are various methods, one of which is based on matching and is commonly known as nearest neighbour imputation (NNI). The NNI method is used in many survey organizations like, Statistics Canada, the U.S. Bureau of Labor Statistics, and the U.S. Census Bureau (Rancourt, Sarndal, and Lee, 1994). When using nearest neighbour for imputation, the missing values are imputed under the assumption that the cases with similar covariates have similar responses.

Matching on covariates allows one to select the respondents with similar covariates to that of nonrespondents thereby reducing the nonresponse bias (Chen and Shao, 2000, Rubin and Rosenbaum, 1983, Zhao, 2004). But when the covariate space is multivariate, it is hard to find matched pairs with the same or even similar values to that of the covariates (X). Even in a simple case when all the variables are binary, there will be 2^p possible values of X where p is the dimension of X, this makes it hard to find matches that are homogenous in X (Rosenbaum, 2002). An alternative in such situations proposed in this paper is to use the propensity matching previously used in a different context by Rosenbaum and Rubin (1983)(RR in the rest of the paper). As defined by RR, a balancing score $b(X)$ is a function of the observed covariates such that the conditional distribution of X given $b(X)$ is the same for missing (nonrespondents, denoted by $m = 1$) and non missing cases (respondents, $m = 0$) and this is denoted by $X \perp\!\!\!\perp m|b(X)$. It is this property that allows the distribution of covariates in respondents and nonrespondents to be similar when matched using covariates or propensity scores, thereby reducing the nonresponse bias. According to RR, matching by covariates provides the finest balancing score, matching on propensity score provides the coarsest balancing score.

Use of propensity scores in missing data was first introduced by Little (1986) for forming strata prior to imputation. Later its use was shown by Lavori (1995) for

*Department of Mathematics and Statistics, University of Canterbury, New Zealand, email: nmi13@student.canterbury.ac.nz

†Department of Mathematics and Statistics, University of Canterbury, New Zealand

multiple imputation by approximate bayesian bootstrap. Both these studies use propensity score for stratifying the data prior to imputation. But in this paper we propose using propensity scores as a method for reducing the dimension of matching variables for imputation. To investigate its efficiency, we compare this method with a NNI method that uses the actual covariates, and a NNI method that uses regression for reducing the dimension of the multivariate space. All the three imputations are carried out on the data which has missingness in one variable.

The rest of the paper is organized as follows. Section 2 describes the nearest neighbour imputation method in general. In section 3 we introduce propensity matching method. Section 4 presents the NNI methods used for comparisons. In section 5 we present the assumptions and details of simulations. The results of the study are given in section 6. We conclude in section 7 that when there are only a few covariates and cases NNI by dissimilarity is the preferred method, but for a large number of covariates, matching by NNI by dissimilarity may be too slow and hence matching by propensity score or regression based nearest neighbour is preferred.

2 Nearest neighbour imputation

To begin with, let us take a simple case and illustrate matching on covariates. Consider a bivariate sample $(X_1,Y_1),(X_2,Y_2).....,(X_n,Y_n)$. Let the variable X be covariate data that is completely observed on all n cases and let Y be observed only for n-r cases. If for any X_j, corresponding to the missing Y_j, we would have an exact match if we can find some X_i corresponding to known Y_i such that $X_i = X_j$. If, as in the general case an exact matching of the covariates corresponding to the missing observations and observations with complete response is infeasible, we weould use the method of nearest neighbour (NN). In Nearest neighbour we match X_j with in the neighbourhood of X. For imputing the missing Y_j, $j = r+1, ...n$, the NN method finds the nearest neighbour using some distance measure. If the distance d_{ij} on the observed X-variables is defined as

$$d_{ij} = |X_i - X_j|, \quad (1)$$

the nearest neighbour obtained for the missing case j is the case k where $d_{kj} = \min_{1\leq i\leq r}(d_{ij})$. In the case where X is multivariate and continuous, one might think of using a distance measure such as Mahalanobis distance. The Mahalanobis distance matching can present problems; for example when a covariate X_i is binary, the Mahalanobis metric may try hard to match this X_i exactly thus reducing the quality of match of the other covariates (Rosenbaum and Rubin, 1983). Another method that is used when the covariate space is multivariate is the NNI method that uses regression (RBNN) (Little and Rubin, 2002). However RBNN methods are not appropriate when the model assumptions are not satisfied (Allison, 2001). In situations where X has different types of variables, Murthy *et-al* (2003) has described a new matching procedure called the nearest neighbour by dissimilarity (NNDM). This takes care of different types of variables, provides efficient matching and preserves the distributions. However when there are many variables, matching on all variables is computationally intensive. Hence, we now propose the use of propensity score for matching. Rosenbaum and Rubin (1983), in the context of observational studies, have shown that propensity matching can effectively balance binary covariates for which matching is not possible on an individual basis. Here we apply this concept to imputation. Since the propensity score effectively represents all variables, we suppose that the use of it for dimension reduction in imputation might be more effective than transforming the multivariate space to univariate space by regression. This has been verified by simulations in section 5.

3 Matching by Propensity score

3.1 Propensity score

The propensity score is defined as follows by Rosenbaum (2002): Let m be the missing indicator defined on Y (the response variable) i.e if Y is observed then $m = 0$ else $m = 1$. The propensity score $\pi(X)$ is defined as $\pi(X)= Pr(m = 1|X)$.

The theory of propensity score given by RR for observational studies and later discussed in context of survey nonresponse by Little (1986) shows that if the missing data are missing at random (MAR) given X then they are missing at random given $\pi(X)$, that is if $m \perp\!\!\!\perp Y|X$ then $m \perp\!\!\!\perp Y|\pi(X)$, where $m \perp\!\!\!\perp Y|X$ means that m is conditionally independent of Y given X (for a proof, see p.48, Little and Rubin, (2002)). In other words, conditioning on the propensity score would remove the correlation between X and m, and hence replacing X with $\pi(X)$ does not lead to any loss of information because $X \perp\!\!\!\perp m|\pi(X)$ (Imbens (2004); Cook (1998)).

The response propensities ($\pi(X)$) are estimated by using logistic regression of m on X. The procedure for computing $\pi(X)$ is described next.

3.2 Computation of propensity score

As in RR we use the following form to estimate the propensity score

$$\pi(X^*) = \frac{e^{\beta X^*}}{1 + e^{\beta X^*}} \quad (2)$$

where $X^* = (1, X_1, X_2, ..., X_p)$ and $\beta = (\beta_0, \beta_1, ...\beta_p)$. The *glm* function in R is used to estimate $\pi(X)$, by regressing m against the covariates X, with the family binomial and link function logit. We do not use the alternate method of discriminant analysis procedure described in Rosenbaum(2002) for estimating the propensity scores because we want a method that deals with a covariate set that has mixed types of variables like binary, nominal, continuous and ordinal variables.

3.3 Nearest neighbour by propensity scores (NNPS)

Having described how we compute the propensity score, we now state some basic ideas behind propensity matching given by Rosenbaum (2002) and then define the distance measure to find the nearest neighbour.

Proposition given in Rosenbaum (2002). *If* $\pi(x) = \pi$, *then*

$$pr[X = x|\pi(X) = \pi, m = 1] =$$

$$pr[X = x|\pi(X) = \pi, m = 0] =$$

$$pr[X = x|\pi(X) = \pi]$$

The above property, known as the balancing property, states that when the propensity score π is the same for the missing and non missing cases then the distribution of covariates is the same for missing and non missing cases. The proposition considered is a particular case (where the number of strata is only one) of the original proposition given in Rosenbaum (2002), hence we avoid the proof.

For NNPS we use equation (1) to determine the nearest neighbour, but with $|X_i - X_j|$ replaced by $|\hat{\pi}(X_i) - \hat{\pi}(X_j)|$, where $\hat{\pi}(X_i)$ is the estimate of $\pi(X_i)$. We use this matching method since the distribution of covariates is the same for missing and non missing cases within the neighbourhood of $\hat{\pi}$. As $\hat{\pi}(X)$ is continuous we expect no ties among the donors obtained for imputing the missing values.

The performance of imputation by propensity matching is next compared to two other nearest neighbour imputation methods in terms of computational time and bias.

4 Methods used for comparison

As explained in section 2, there are several methods for finding the distance in the multivariate case. In this paper we use two procedures; nearest neighbour based on regression (RBNN) and nearest neighbour by dissimilarity (NNDM). These two procedures are described in the following two subsections.

4.1 Regression Based Nearest Neighbour (RBNN)

The RBNN procedure was initially given by Laaksonen (2000). RBNN is similar to the predictive mean matching method given in Rubin (1987). Under this procedure imputation is carried out in the following manner where we use the subscripts obs (mis) to refer the observed (missing) cases;

1. Using the observed cases construct a regression model
$$y_{obs} = \alpha + \beta X_{obs}$$
2. Use the estimates $\hat{\alpha}, \hat{\beta}$ of α, β to predict y (by $\hat{y}$) for all available X
$$\hat{y} = (\hat{y}_{obs}, \hat{y}_{mis}) = \hat{\alpha} + \hat{\beta}X + \epsilon$$
where ϵ is as defined below.
3. Find nearest neighbour to $\hat{y}_{mis}$ from $\hat{y}_{obs}$ for each missing case.
4. Use the y_{obs} corresponding to the nearest $\hat{y}_{obs}$ as the imputed value for y_{mis}

The error (ϵ) is assumed normally distributed (0,σ^2). It is not uncommon to assume $\sigma^2 = 0$, but this does not add variability due to imputation. If the data sets are small and the error term is added it make a difference to the final data (Allison, 2001). There are several ways to add this error term to the model (Kalton and Kaspryzk, 1987), but we use the method of Laaksonen (2000) where σ^2 is the residual mean square error estimated from the regression model.

4.2 Nearest neighbour by dissimilarity matrix (NNDM)

NNDM is the other method used for comparison as it provides good matching when the variables are of mixed type and it preserves the distributions as shown in Murthy *et-al* (2003).
The dissimilarity matrix between a complete case c and a missing case m is defined as

$$D(c,m) = \frac{\sum_{j=1}^{p-1} \delta_{cm}^{j} d_{cm}^{j}}{\sum_{j=1}^{p-1} \delta_{cm}^{j}} \quad (3)$$

where the distance d_{cm} for the j^{th} covariate (d_{cm}^{j}) is given by

$$d_{cm}^{j} = \begin{cases} \left.\begin{array}{ll} 1 & \text{if } x_{cj} \neq x_{mj} \\ 0 & \text{otherwise} \end{array}\right\} \text{for binary nominal variables.} \\ \frac{|x_{cj} - x_{mj}|}{r_j} \quad \text{for interval and ordinal variables} \end{cases}$$

δ_{cm}^{j} is an indicator variable which is 1 except when the j^{th} covariate is asymmetric and $x_{cj} = x_{mj} = 0$, and r_j is the range of the j^{th} covariate.
We use r_j rather than standard deviation to normalize as this ensures that for interval and ordinal variables $d_{cm}^{j} \in$ [0,1] as for the other binary and nominal variables. For further details on asymmetric variables refer to Kaufman and Rousseeuw (1990) or Murthy*et-al* (2003).
Of possible donors the case c with $min[D(c,m)]$ is considered the nearest neighbour and used as the donor. If there are several donors with the same $min[D(c,m)]$, a donor may be randomly selected from among them.

5 Simulation

For comparing the closeness of the imputed values by proposed matching methods, we use Monte Carlo simulations. In these simulations we assume that; a) Missingness is in one variable, b) the sample is drawn by simple random sampling, c) there is only one imputation class and, d) the missing data is MAR linear or MAR convex. The Monte Carlo simulations were performed on three different data sets.

5.1 Data generation

We used two real life data sets and a simulated data set for these comparisons. The two real life data sets used are; Tooth Growth data (TGD) (Mc.Neil,1977) and Low birth Weight data (LBW) (Hosmer and Lemeshaw, 2002). As these data sets had sample sizes of 60 and 189 respectively, to investigate the performance of these methods on large data sets, we simulated an artificial population with the covariate (X) being multivariate. In this artificial data set (AD), we have three randomly generated covariates X', Z, W, of size $N = 10,000$ and a response variable Y. In the artificial data, the variable Z is a binary variable and W is a categorical variable with three categories. The covariate X' and the response variable Y have a joint distribution $N(\mu, \Sigma)$, where $\mu = (10, 12)$ and $\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 2.5 \end{bmatrix}$. All the three data sets initially have no missing values and missingness is induced into the data using MAR linear and MAR convex mechanisms. These mechanisms are detailed in the following subsection.

5.2 Missing data mechanism

The different types of missingness used in comparisons are MAR linear, and MAR convex as defined in Collins *et-al* (2001). The MAR mechanism is

1. MAR linear if the probability of missingness is linearly related to one of the covariates.

2. MAR convex if missingness is more on the extremes of the covariate and smaller in the middle

For MAR linear, missingness is created by dividing a continuous covariate into four groups based on quartiles and then set different probabilities in a linear manner such that we achieve a desired percent of missingness. For MAR convex missingness is created by first dividing a covariate into four groups based on the quartiles. The probabilities of nonresponse are set high in the first and last quartile and low in middle two quartiles to achieve a desired amount of missingness. In these simulations a covariate which is of interval type is used to form the groups.

5.3 Creating missing data

In all the three data sets missingness is induced using MAR linear and MAR convex mechanism as described in sec 5.2.
Missing data in artificially generated data (AD): For each simulation a sample of size 1,000 is drawn for the population of size 10,000 using simple random sample with replacement. We used sample with replacement in order to be consistent with the sample selection process of

the tooth growth and low birth weight data simulations where, because of the small data length, we took sample with replacement. We call this sample data set as "ADS". In order to achieve 25 and 40 percent missing rates in ADS we set the probabilities to (0.1,0.2,0.3,0.4) and twice these values under MAR linear. Similarly we set the probabilities to (0.4,0.1,0.1,0.4) and twice these values to achieve 25 and 40 percent missing data under MAR convex mechanism. X' variable of AD was used to form the quartiles.

Missing data in tooth growth data (TGD): The data set TG has three variables and is of size 60. the three variables are "length", "dose", and "supplement" (supp). Variables length and dose are of interval type and supp is binary. Length and dose are correlated the correlation being 0.80. Instead of using an interval variable to form quartiles we used a binary variable by doing so we created a special case where MAR linear and MAR convex are the same. Now in this data the probabilities were set to (0.2 and 0.4) to have around 28 percent missing data in all the simulations.

Missing data in low birth weight (LBW): This data set has an asymmetric binary variable (Smoking). The variable Age was used for generating the missingness. The probabilities for MAR linear were (0.1,0.2,0.3,0.4) for 18% missingness and (0.2,0.4,0.6,0.8) for 35 % missingness. For MAR convex the probabilities were (0.4,0.1,0.1,0.4) and (0.8,0.2,0.2,0.8) to achieve 18 and 35 percent missingness. 1,000 simulations were carried out. We used 9 out of 11 variables in our simulations omitting the variables "Case ID" and "Low". The variable "Low" in the LBW data is a categorical variable constructed from birth weight information. As we are using birth weight as a continuous variable, the variable Low is redundant. This data set has 189 observations.

In order to compare the performance of the procedures we used the mean square errors (MSE). The mean square error is computed as the squared difference between the original value before missingness is inserted and the imputed value; that is

$$\Delta = \frac{\sum_{i=1}^{n_{imp}} (y_i^{imp} - y_i^{actual})^2}{n_{imp}}$$

where y^{imp} is the imputed value and y^{actual} is the actual value of y before inducing the missingness and n_{imp} is the number of imputed values.

6 Results

Simulation Set 1:
For the first set of simulations we used ADS data. For the RBNN method the regression model had an R^2 value in between 0.43-0.79. The error term ϵ is generated with normal distribution $(0, \sigma^2)$, where σ^2 is the residual mean square error obtained from the regression as defined in sec.4.1.

The imputation results presented in Tables-1, show that NNDM imputes the missing values close to the true values in all cases. Comparison of NNPS and RBNN shows that RBNN has lower MSE under MAR convex. For MAR linear the NNPS is the preferred choice.

Simulation Set 2:
In this simulation set we used the tooth growth data. For this data the details of the regression model used in RBNN are; R^2 lies in the interval 0.65-0.72.

Imputation results are presented in Table-2. For this data the NNPS and NNDM imputes the missing values close to the true values.

Simulation set 3:
For the third set of simulations we used LBW data. Results of the regression model used in RBNN show that R^2 lies in between 0.2 and 0.45. From Table 3 it is observed once again that NNPS may be a better option than RBNN. The MSE of NNPS lies in between that of RBNN and NNDM. The MSE presented in Tables 1,2, and 3 is the mean of the MSE's obtained from 1000 simulations. The standard error reported in the tables is obtained using the bootstrap function in the R package.

Computational Time:
Table-4 presents the computational times for all the methods. The comparisons of the computational times show that there is a notable reduction in time when imputation is performed using NNPS or RBNN. For the TGD simulation, the computational time for a single simulation run by NNDM is 0.26 seconds and by RBNN and NNPS 0.03 seconds. When the number of cases were increased to 1,000 as in the ADS data, the computational time for NNDM is 29.55 seconds for MAR linear and 25% missingness, when the percent missing is around 40 the computation time increased to 73.6 seconds. For RBNN and NNPS the computational times are 0.34 and 0.50 with 25% MAR linear missingness and 0.56 and 0.88 with 40% missingness. MAR convex also gave similar results.

Table-1: Comparison of MSE under MAR linear and MAR convex with different % of missingness in simulated data.

% Missing	MAR	Matching by	MSE	SE
25	Convex	NNDM	10.91	0.012
		NNPS	11.46	0.013
		RBNN	11.00	0.013
	Linear	NNDM	10.85	0.008
		NNPS	10.96	0.009
		RBNN	11.08	0.010
40	Convex	NNDM	10.96	0.011
		NNPS	11.57	0.010
		RBNN	10.99	0.010
	Linear	NNDM	10.86	0.008
		NNPS	10.87	0.009
		RBNN	11.01	0.010

Table-2: Comparison of MSE for the Tooth Growth data with 28 percent missing data.

Matching by	Missing Mechanism	
	MAR	
	MSE	SE
NNDM	6.39	0.002
NNPS	7.06	0.004
RBNN	15.27	0.006

Table-3: Comparison of MSE under MAR linear and MAR convex with different % of missingness in Low birth weight (LBW) data.

% Missing	MAR	Matching by	MSE	SE
18	Convex	NNDM	0.334	0.002
		NNPS	0.38	0.005
		RBNN	0.38	0.005
	Linear	NNDM	0.22	0.003
		NNPS	0.24	0.005
		RBNN	0.25	0.004
35	Convex	NNDM	0.33	0.003
		NNPS	0.36	0.006
		RBNN	0.40	0.006
	Linear	NNDM	0.34	0.003
		NNPS	0.40	0.004
		RBNN	0.46	0.009

Table-4 Computational time in seconds for the three methods under different percent of missingness.

Data set	% missing	Matching by		
		NNDM	NNPS	RBNN
AD	25	29.5	0.34	0.50
	40	73.6	0.56	0.88
LBW (linear)	18	1.53	0.03	0.02
	35	1.84	0.05	0.03
LBW (convex)	18	0.84	0.03	0.02
	35	1.42	0.03	0.02
TGD	28	0.26	0.03	0.02

7 Conclusions

From the results obtained, we observed that, when there are few covariates and cases the use of all covariates to find nearest neighbour is recommended; when the covariates are of different types NNDM should be used. When the number of covariates and/or the number of cases increase, NNDM is still more accurate but may be too slow. In these cases the NNPS method is better if the missingness in data is MAR linear and the RBNN method is better if the missingness in data is MAR convex. However since this is based on a limited number of situations, further study will be needed to confirm these findings. In particular we intend to apply these methods to the National Family Health Survey data and investigate their effectiveness.

References

Allison, P.D. (2001). *Missing Data.* Series: quantitative applications in the social sciences. Sage Publications: Thousand Oaks.

Chen, J., and Shao, J. (2000). Nearest neighbour imputation for survey data. *Journal of official statistics*: Vol 16, No. 2, pp 113-131.

Collins, L.M., Schafer, J.L., and Kam, C.M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*: Vol 6, No.4, pp 330-351.

Cook, R.D. (1998). *Regression Graphics: Ideas for studying regression through graphics*, John Wiley: New York.

Hosmer, D.W., and Lemeshaw, S. (2000). *Applied logistic regression.* 2^{nd} edition. John Wiley: New York.

Imbens, G.W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review, *The Review of Economics and Statistics*: Vol 86, No.1, pp 4-29.

Kalton, G., and Kaspryzk, D. (1981). Imputing for survey response. *Proceedings of the section on survey research methods.* American statistical association. pp 22-33.

Kaufman, L., and Rousseeuw, P. J. (1990). *Finding Groups in Data. An introduction to cluster analysis.* John Wiley: New York.

Laaksonen, S. (2000). Regression based nearest neighbour hot decking, *Computational statistics*: Vol 15, No.1, pp 65-71

Lavori, P.W., Dawson, R., and Shera, D. (1995). A Multiple imputation strategy for clinical trials with truncation patient data, *Statistics in Medicine*: Vol 14, pp 1913-1925.

Little, R.J.A.(1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*: Vol 54, pp 139-157.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing data.* 2^{nd} edition. John Wiley: New York.

Marker, D.A., Judkins, D.R., and Winglee, M. (2002). Large scale imputation for complex surveys. *Survey Nonresponse.* Ed. Groves, R.M., Dillman, D.A., Eltinge, J.l., and Little, R.J.A. John Wiley: New York. item Mc.Neil, D.R. (1977). *Interactive Data Analysis* Wiley: New York.

Murthy, M.N., Chacko, E., Penny, R., and Monir Hossain, Md. (2003). Multivariate nearest neighbour imputation. *Journal of Statistics in Transition*: Vol 6 , No.1 , pp 55-66

Rancourt, E., Sarndal, C., and Hyunshik Lee. (1994). Estimation of the variance in the presence of nearest neighbour imputation.*Proceedings of the section on survey research methods.* American Statistical Association. pp 888-893.

Rosenbaum, P.R. (2002). *Observational studies.* 2^{nd} edition. Springer-Verlag: New York.

Rosenbaum, P.R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*: 70, No. 1, pp 41-55.[Denoted as RR in the paper]

Zhao, Zhong. (2004). Using matching to estimate treatment effects data requirements, matching metrics, and monte carlo evidence, *The Review of Economics and Statistics*: Vol 86, No.1, pp 91-107.

Acknowledgements

The authors wish to thank Richard Penny, and Jana Asher for their comments and help in improving this paper. Earlier version of this paper presented at JSM 2004,Toronto.

Bootstrap Variance Estimation for Two Stage Cluster designs with Imputed Data

Murthy. M. N., Chacko Easaw
University of Canterbury, Department of Mathematics and Statistics, Christchurch, New Zealand
E-mail Address nmi13@student.canterbury.ac.nz

Richard Penny
Fidelio Consultancy, 264 Grahams Road, Christchurch, New Zealand

Imputation is a commonly used tool for adjusting the item nonresponse in large scale surveys. Treating imputed data as true observations and using naive variance estimation would underestimate the variance. As a way of accounting for the additional variance due to imputation, Rubin introduced multiple imputation. For single imputation, Sarndal and several others since then have looked at variance estimation for the imputed data. Shao and Sitter (1996) and others used resampling methods. In this paper we adapt the bootstrap method to find the variance of the mean obtained from an imputed data under two stage cluster designs. As is common we assume here that missingness is in one variable (Y) and there is at least one complete covariate variable (X) in the sample.

The Bootstrap for imputed data

Shao and Sitter (1996) have described a bootstrap procedure for finding the variance of the statistic of interest obtained from data imputed by a simple random hot deck method both in case of simple random samples (SRS) and stratified designs. However even for the case of SRS, the method suggested by Shao and Sitter (SS), has limitations. When the data has low nonresponse rate and the sample size is small, the bootstrap samples may not contain any nonresponse and hence might not be representative of the original sample. When the data has more than 50 percent nonresponse, there may be too many repetitions of the same donors thus creating a spike. In order to overcome this situation Bello (1994) suggests that it is important to have the proportion of nonresponse in the bootstrap sample to be the same as in the original sample. With this in mind, we make a modification to the bootstrap. The modified bootstrap method is first applied to a simple random sample and then to two stage cluster designs.

Modified Bootstrap

First consider the case of SRS. In a sample of $n = r + m$, let Y_{obs} be the r observed cases and Y_{mis} be the m missing cases. Let Y^*_{mis} be the imputed values for the m missing cases. Denote the imputed data set by $Y_I = (Y_{obs}, Y^*_{mis})$. Let R be the indicator variable, with $R = 1$ if Y is observed. In this paper we use nearest neighbour method for imputing the missing values. Our modified procedure is as follows:

1. Compute the propensity score using $Z = (X, R)$
2. From the n cases we draw with replacement $n^* = n$ cases in the bootstrap samples.
3. Choose m of these propensity scores with the lowest, and assume that they are nonrespondents.
4. The remaining r cases will all or mostly be respondents but if there are any nonrespondents, replace them with randomly selected respondents from the original set of respondents.

5. Impute the missing data with the same imputation procedure that is used for imputing the missing data in Y_I.

6. Compute the statistic $\hat{\theta}_b^*$ of interest from the imputed b^{th} bootstrap sample (Y_I^{*b}).

7. Repeat steps 2-6 for $b = 1, 2, ...B$ times and compute the variance estimate by

$$(1) \qquad s_B^2(\hat{\theta}) = \sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\hat{\theta}}^*)/(B-1) \text{ where } \bar{\hat{\theta}}^* = \sum_{b=1}^{B}(\hat{\theta}_b^*)/B$$

Selecting the bootstrap sample in this manner would ensure that each bootstrap sample has the same proportion of missingness in every draw. We make use of this modified bootstrap method for finding the variance of the imputed data in two stage cluster design.

Extension of The Bootstrap Method to Cluster Designs

The simple bootstrap when applied to the case of full response in two stage cluster designs would underestimate the variance (Rao and Wu, 1988), hence they suggest a correction. For imputed data, in this paper we propose a similar adjustment when the design is two stage cluster. Our extension assumes that imputation is done within a cluster. In addition we also assume that the initial sample data (both, clusters and 2nd stage samples) was drawn using simple random sample without replacement. Here we assume that there are k clusters in the initial sample with the cluster i having a sample of size n_i. The bootstrap method is as follows:

1. From these k clusters we randomly sample $k^* = k$ clusters with replacement.

2. Use steps 2-5 of the modified bootstrap above for each selected cluster.

3. Use the adjustment procedure in Rao and Wu (1988, eq 6.4, p.239) to get $\hat{\theta}_b^*$ for the b^{th} bootstrap.

4. Repeat steps 1-3 for $b = 1, 2, ...B$ times and then,

5. Find the variance of the statistic using eq (1).

Preliminary results indicates that this modified bootstrap method has promise and, for the data sets tried gives variance estimate that lies between the true variance and that given in SS method.

REFERENCES

Bello, A.L. (1994) A bootstrap method for using imputation techniques for data with missing values. *Biometrical Journal*, 36 (4), 453-464.

Rao, J. N. K. and Wu, C. F. J. (1988) Resampling inference with complex survey data, *Journal of American Statistical Association*, 83 231-241.

Shao, J. and Sitter, R. R. (1996), Bootstrap for imputed survey data, *Journal of the American Statistical Association*, 91, pp.1278-1288.

RÉSUMÉ

On propose une méthode 'bootstrap' modifiée pour éstimer la variance des données imputées. Pour les échantillons 'bootstrap', cette méthode maintient la proportion de manque des données originales. On étend cette méthode aux cluster designs de deux étapes.